論文内容の要旨

論文題目　Computer analysis of DNA sequence patterns that define transcription, splicing, and protein coding in human genome
転写、スプライシング、タンパク質コード領域を決定するヒトゲノム DNA 配列パターンのコンピューター解析

氏名　村上勝彦

ヒトのゲノム配列と転写産物のデータが蓄積されてきた現在、ヒトの DNA 配列を通して生物学的現象や機能を解析する能力は飛躍的に高まった。転写産物をゲノムにマップした後、どのような配列が何の生物学的機能を制御しているかを解明するということは、重要かつチャレンジングな問題である。本論文の主目的は、生物学的機能を持つ DNA 配列の特徴を見出すことである。もう一つの目的は、DNA 配列を主としてみたときに、生物学的特徴を計算機科学的方法でどの程度説明できるのかを調べることである。この目的を達成するために本論文では、ヒトゲノム配列中で既知の機能領域（タンパク質コード領域、スプライス部位およびプロモーター領域）を用いて、機能と配列の関係を網羅的に解析した。

本論文の構成は大きく２つに分けられる。前半では、ゲノム配列からの遺伝子構造予測について述べている。ここでは、「複数プログラムの統合による遺伝子予測」と、「スプライス部位のクラスタリング」を行った。後半では、ヒト遺伝子のプロモーターの構造解析について述べている。ここでは、「（予測された）転写因子結合部位（モチーフ）のクラスターの存在とプロモーターの相関解析」と、「モチーフペアと転写開始点（TSS）の位置関係の解析」を行った。

初めに申請者は複数の遺伝子予測プログラムの予測結果を統合することにより予測精度を向上させるということを提案した。転写産物データが大量に蓄積された現代でも、遺伝子の発現頻度が低い、または稀な組織でしか発現しない、などの理由によって、実験で同定が困難な、未知のタンパク質コード遺伝子が、まだ存在すると思われ、それらを効率的に見出すために予測精度を上げることは重要である。申請者は幾つかの統合方法を評価した上で、異なるプログラムの結果を比較可能にする改良スコアを計算し、これが高い予測候補を選別する方法が約５％の精度向上をもたらすことを示した。

次に、遺伝子予測精度向上の要となる、５’スプライス部位をゲノム配列から識別する際の精度向上を目的に、スプライス部位内の位置塩基依存性を考慮した新規クラスタリング方法および識別方法を提案した。この方法によって、スプライス部位は５つのクラスのいずれかに属するとして認識されるが、特定のクラスに分類された場合には、それまでの方法に比べて、擬陽性を減らして認識できることを示した。この結果を上手く統合すれば、遺伝子構造全体の識別精度向上が期待できる。向上したクラスのうち、あるクラスの配列は、５’スプライス部位に作用する U1snRNP の配列に近いものになっていたが、このことから、本結果の妥当性と、各クラスで異なる認識メカニズムの存在が示唆される。

後半では、ヒト遺伝子のプロモーターの構造解析について記述している。プロモーター領域は、遺伝子転写開始点付近の領域およびその上流領域であり、そこに基本転写因子や様々な調節因子といった転写因子が結合部位（モチーフ配列）に結合して転写開始複合体を形成し、遺伝子の転写を行う。従って、プロモーター領域のモチーフ配列が、転写制御を解明する鍵である。どのようなプロモーター領域であれば、どのような転写制御を受けるのかという問題は重要であるが、現在では個別に実験で解析した事実を積み上げている段階で、まだ統一的な理解や記述ができているとはいえない。現在、脊椎動物一般のプロモーター領域では、複数のモチーフ配列が集まってモジュールを構成しているという仮説が有力視されている。さらに、その

中には同一のモチーフが多いケースが散見される。転写を促進させるためには、必要な転写因子がプロモーター領域に集まらなければならない。もしプロモーター領域に同一モチーフのクラスターが存在すると、転写因子がその領域に滞留する可能性や、転写促進効果のより高い近傍のモチーフと相互作用をする可能性が高くなり、転写の活性化に役立つと思われる。この仮説を元にプロモーター領域予測をするツールも開発されていた。そこで申請者は、プロモーター領域には同一モチーフクラスターが存在するという仮説が、モチーフによらず一般的に成り立つかどうかを調べるため、同一モチーフのクラスターとプロモーター領域との相関をモチーフ毎に調べた。その結果、約４７％がモチーフクラスターとプロモーターに相関があり、仮説が一般には成り立たずモチーフに依存して成り立つことを明らかにした。このモチーフクラスターとプロモーターの相関には、CpG アイランドとの関連が考えられる。そこで、CpG アイランドと関連して相関があるのか、あるいは独立に相関があるのかを検討した。３者関係から２者間のみの相関を取り出すために偏相関係数を用いた。その結果、CpG アイランドと関連して相関があるモチーフ群（１３％）と、CpG アイランドと独立に相関があるモチーフ群（２３％個）を同定した。この２つのモチーフ郡を用いて、遺伝子の組織特異性の予測指標を提案した。結果として単純には予測は困難であるが、有意に差の出る結果を得た。

次に、プロモーターの更に複雑な構造を解明するため、モチーフがペアで同じプロモーター領域に（重ならない位置関係で）共起する場合に、転写開始点（TSS）からの相対位置にバイアスが有意に存在するかどうかを調べた。ここでは異なるモチーフの共起も、同じモチーフの共起も対象としている。これまでに、モチーフ単独でも TSS からの位置特異的に頻出する（頻度分布でピークが見られる）ものが知られていた（TATA-box, SP1 モチーフ等）。申請者は、モチーフペアが共起する位置のバイアスを計算する指標を設け、χ 二乗検定および補足的な統計検定によって有意となるペアを検出した。その結果 8,928 個(34%)のペアが FDR<1%の基準で有意に検出された。位置のバイアスをヒートマップにより可視化すると様々なパターンが見られたが、ここでは、TSS 付近でペアのピークがあるもの（クラス１：２５０ペア）と、ペア間距離が２００ｂｐ以下で且つ TSS からの距離に関係ない位置でピークがあるもの（クラス２：４，１９９ペア）およびその他（クラス３：４，４７９ペア）に分類した。クラス１に頻度の高いモチーフは、モチーフ単独でもピークを持つモチーフであったが、一方、クラス２のそれは、単独ではピークがもともと見られないモチーフであって、２クラスの主たる構成メンバーは有意に異なっていた。各クラスを上流に持つ遺伝子について、その発現の組織特異性を EST のユニークライブラリ数で調べると、クラス１ではより広い組織に発現が見られた。次に、ペアの中で頻度の高いモチーフに結合する転写因子のタンパク質ドメインを全クラスで調べたところ、特定のアミノ酸がリッチなドメインが多く、そのアミノ酸群はタンパク質間相互作用で重要なディスオーダー領域の特徴と重なった。これはモチーフペアに結合する転写因子が相互作用をする際に、ディスオーダー領域が関係するような特定の相互作用をしている可能性を示唆するものである。このドメインの偏りをクラス毎に見た場合、クラス１よりクラス２で非常に多かった。これらから申請者は、polII 系で転写される全ヒト遺伝子のプロモーター構造の特徴として、共起する位置に有意なバイアスがあるようなモチーフペアが数多く存在していることを示し、それらのペアには特徴の異なる２つのクラスが存在するという見方を提案した。

以上のように、申請者はゲノム配列とそこに内在する生物学的機能を関連づける一連の手法を提案し、種々の知見を得た。これらはヒトゲノム配列をより深く理解する上で有用なものである。