JAIST Repository
https://dspace.jaist.ac.jp/

Title	遺伝子転写制御領域に含まれる特異的文字列の解析とDNAマイクロアレイデータを用いた遺伝子間の依存関係推定
Author(s)	上田，智之
Citation	
Issue Date	2004-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1805
Rights	
Description	Supervisor:平石 邦彦，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

遺伝子転写制御領域に含まれる特異的文字列の解析とＤＮＡマイクロアレイデータを用いた遺伝子間の依存関係推定

上田 智之 (210006)

北陸先端科学技術大学院大学 情報科学研究科

2004年2月13日

キーワード: 遺伝子間の依存関係、特異的文字列、統計的解析、類似性.

ゲノムの語源は遺伝子（gene）＋染色体（chromosome）で、ゲノムという言葉は、染色体上の遺伝子が持つ情報を意味する。現在、ゲノム研究の対象は遺伝子の構造的な情報から機能的な情報の解明へと移りつつある。遺伝子の発現によって生成される蛋白質の生成過程においては、様々な要因が蛋白質の生成に影響を与える。特に、ある遺伝子Ａから生成された蛋白質がある遺伝子Ｂの転写に影響を与えるとき、転写を制御する蛋白質を転写制御因子と呼び、遺伝子Ａと遺伝子Ｂの間には依存関係があるという。つまり、遺伝子間の依存関係とは遺伝子Ａから生成された蛋白質が遺伝子Ｂの転写制御領域に特異的に結合し、遺伝子Ｂの転写を制御することによって、遺伝子Ｂの蛋白質生成を制御することをいう。また、このとき、遺伝子Ａを調節遺伝子、遺伝子Ｂを被調節遺伝子と呼ぶ。調節遺伝子と依存関係にある遺伝子の転写制御領域には、塩基配列（文字列）の類似性があることが知られている。また、その類似配列を表した特異的文字列パターンをもつ遺伝子は共通の制御を受けることが統計的に有意であり、転写制御領域の解析は依存関係推定に有効であるといえる。また、遺伝子間の依存関係推定における従来研究には、ブーリアンネットワークやベイジアンネットワーク、Ｓ－ＳＹＳＴＥＭがある。これらはＤＮＡマイクロアレイデータのみを用いた手法がある。ＤＮＡマイクロアレイデータとは、同時に多くの遺伝子の発現を観測したデータである。しかし、データの信頼性が低いことや複数の準最適解があることなどの問題がある。このような理由から、ＤＮＡマイクロアレイデータのみではなく、蛋白質間相互作用や蛋白質－ＤＮＡ相互作用、プロモーター領域に含まれる共通配列といった別の生物学的情報を付加した推定方法が注目されている。プロモーターシーケンスに関する研究には、コンセンサスシーケンスやモチーフなどがある。また、近年、様々な生物のＤＮＡ塩基配列が決定されたことにより、構造的な情報を用いた研究が可能となっている。本研究では既に全塩基配列や遺伝子の位置情報が決定されている枯草菌を用いる。従って、本研究では遺伝子転写制御領域に含まれる特異的文字列の解析とＤ

Copyright © 2004 by Ueda Tomoyuki

ＮＡマイクロアレイデータを用いて、遺伝子間の制御関係を推定する。本研究の手法は、まずはじめに、遺伝子転写制御領域の統計的な解析によって特異的文字列の候補を取り出す。次に、各遺伝子転写制御領域に含まれる特異的文字列を対象として、局所的な類似性を評価し、その最大値を遺伝子間の類似度とする。そして、類似度の高い遺伝子群とＤＮＡマイクロアレイデータの発現強度に強い相関がある遺伝子群の両方に含まれる遺伝子群を制御関係があると推定する。調査の準備として、枯草菌の全遺伝子中の転写制御領域の長さが $0[bp]$ 以上の遺伝子を対象とし、ＮＢＣＩ（http//www.nbci.nlm.nih.gov）で公開されている遺伝子の開始位置を転写制御領域の開始位置として、それより上流 $529[bp]$ までを各遺伝子の転写制御領域として抜き出す。次に、抜き出した各遺伝子の転写制御領域に含まれる色々な長さ（6〜10）の文字列の出現頻度を調べた。結果、長い文字列では出現頻度 1 がである文字列が多く、特異的文字列の候補となる文字列が多い。また、長い文字列を部分文字列によって表現することができる。従って、以後の調査では長さ 6 の文字列を対象とし、文字列の統計的特異性を定量化するために、各文字列 s の出現頻度を O_s 、期待値を E_s としたとき、特異度を $(O_s - E_s)/E_s$ により定義した。そして、この特異度によって表された転写制御領域では既知の特異的文字列の多くが特異度がマイナスの領域に含まれていることが発見された。さらに以後の調査で、特異度（ -0.1 ）以下では既知の特異的文字列の $90[\%]$ をカバーすることが明らかとなり、その閾値以下の文字列（2294 種類の文字列）を特異的文字列の候補とした。遺伝子間の類似性評価では、このような特異性の高い文字列を対象として、各遺伝子対に対し、転写制御領域のウィンドウ（長さ 30 文字の連続領域）に含まれる特異的文字列の候補どうしの類似性を計算し、その最大値を２つの遺伝子間の類似度とした。そして、この手法を用いて、ある遺伝子についての調査を行ない、得られた類似度の中で高いスコアの遺伝子群とＤＮＡマイクロアレイデータの発現強度に相関がある遺伝子群の両方に含まれる遺伝子群を制御関係があると推定する。これを依存関係既知の遺伝子群に対して適用してみた結果、phoP から影響を受ける遺伝子群では平均スコアよりも高い値を示す遺伝子が多くみられた。しかし、そのようなスコアを示す領域は多くの遺伝子に影響を与えるσ因子が結合する場所であった。また、cqpA から影響を受ける遺伝子群では個々の遺伝子間のスコアは一様に分布しているように観測されたが、調査した遺伝子群のトータルスコアを平均スコアと比較した場合、高いスコアを示す遺伝子が多かった。これは、個々の遺伝子の比較においてスコアのバラつきが小さいためと考えられる。また、生物学的な研究により、転写制御因子は同じ一次元構造であっても、立体構造の違いを認識して結合することが知られている。したがって、今後の調査では既知の遺伝子群において各遺伝子間のスコアと位置、トータルスコア等の評価方法及び最適な条件を調査し、また、転写制御因子が持つ立体構造の認識能力を考慮して、遺伝子間の制御関係を類似度と DNA マイクロアレイデータを用いて推定する。