P5

統計的コンセンサス配列及び実験データによる遺伝子転写開始点上流配列の統合的特徴づけ

Integrated Characterization of Transcriptional Upstream Sequences by Statistical Consensus Patterns and Experimental Data

藤渕　航、金久　實（京都大学化学研究所）
Wataru Fujibuchi and Minoru Kanehisa
Institute for Chemical Research, Kyoto University

Abstract

Transcription is one of the key phenomena for gene expression, and it is important to clear the mechanism of transcriptional control on DNA sequences. But there are so many concerning parameters which should be taken into account to clarify the mechanism that it is difficult to handle them at once with experimental method. To overcome this problem, we have been developing a computer system which deals with much information for transcription.

We first collected promoter sequences from EMBL release 31 database, and aligned them with the transcriptional initiation site. To extract consensus patterns with locational constraints, we calculated appearance frequencies for all patterns on each window position of all sequences. Allowing base substitution, we put similar patterns together and filtered them with threshold to construct a consensus pattern index.

In addition to consensus information, we used TFD(Transcription Factor Database) and EPD(Eukaryotic Promoter Database) as experimental data for identifying binding sites of transcription factors and the type of RNA polymerase, respectively.

With these statistical and experimental data, we retrieved promoter sequences and represented them as a set of significant information. The representation is not only useful for suggesting a lot of considerable parameters but will be possible to search similar promoter by functional aspect.

1　はじめに

ゲノム計画の進行に伴い、DNAの塩基配列から遺伝子が同定されてくると、次に期待されるのはそれら個々の遺伝子がどういった制御を受けて発現し、蛋白質に変換されているのかという、総合レベルでの理解である。遺伝子の発現制御を理解する上で、転写による調節は重要な位置を占めている。しかしながら、DNA配列上で転写に関する情報は様々なシグナルの形で埋めこまれており、個々のパラメータの総合的結果として転写が起こるか否かが決定されることが知られている。これらのパラメータを全て念頭に置いて、実験を行うことは非常に困難であり、よって計算機により転写に関するパラメータをなるべく多く表現させ、それらを統合的に処理しながら実験を進めて行く方法が役立つと考えられる。

そこで我々は、統計的手法によって得られたデータと実験的に得られているデータ

を使い、統合的に遺伝子の転写上流域を特徴づけてやり、生物学的解析を可能とするような計算機上のシステムを開発してきた。今回は統計的手法としては、位置情報を取り入れたコンセンサスパターンの検出に限り、実験データとしては、転写因子の結合サイト情報とRNA polymerase II系の転写と同定されているかの情報の2つを用いている。

2　統計的手法によるコンセンサスパターンの検出

位置情報を取り入れたコンセンサスパターンは次のようにして抽出してきた。

EMBL(release 31) の features table からprimary transcript データを取りだし、転写開始点上流の配列を収集する。これらから、必要とする配列長を満たさないものや、ベクターのコンタミネーション配列のあるものを除き、更に配列間で総当たりのアライメントを取ることでホモロジーの高いデータは除いておく。

転写には alternative transcription という現象が見られるので、データベースからalternative な転写開始点の相対的距離情報を抽出し、後にウインドーサーチ[1]を行うときのウインドー幅決定への参考とした。

位置情報を入れて配列パターンを得るため、これらの配列を転写開始点でアラインしておいたものを、各塩基位置でウインドーサーチをかける。今回は、パターンの取りうる自由度と計算機のメモリ上の問題から6塩基のパターンをウインドーあたり１１回サーチし、中央のパターンの先頭塩基位置に記録していった。但し、この時同じウインドー内では同一パターンはその出現回数を一度しか数えない。この all /none による数え方は、同一ウインドー内で何度も同じパターンを数える従来の方法に比べ、データの歪みを拾いにくいことが示された。

拾ってきたパターンをグループ化するため、塩基の置換を許して類似パターンを集約してゆく。この時、塩基の置換を許し過ぎるとそのパターンのもつあいまいさが増えるので、塩基置換をしていないときのパターン間の相対頻度の差がそれほど崩れないまでの1塩基置換にとどめている。

次に来るのは有意パターンのセレクションの問題である。その位置で最も高頻度で出現した6文字パターンが $a_1a_2a_3a_4a_5a_6$ であったとすると、普通それを一文字右や左にシフトしたパターンである $xa_1a_2a_3a_4a_5$ や $a_1a_2a_3a_4a_5x$ も高頻度で取れてくる。今回これらのシフトパターンはその中央のもののみが価値があるとし、3文字シフトしたパターンまでは除去した。

最後に、それらをしきい値（出現頻度＞データ数の10%）を用いてコンセンサス度の高いもののみを残し、各パターンが配列上でどのくらいの出現位置の幅をもっているかをまとめ、出現開始位置、終了位置、パターン、頻度を表記したコンセンサスパターンインデックスを作成した(Fig. 1)。

3　転写に関する実験データの利用

転写に関する実験側からのデータとして、現在 TFD(Transcription Factor Database)[2] と

EPD(Eukaryotic Promoter Database)[3] が知られている。TFD(release 5.0) の方にはこれまでに見つかっている転写因子とその結合サイトに関する情報が記載されており、これから unknown factor と表記されているものを除いた1547種の「転写因子名－結合サイト」の情報を特徴づけに使用した。またEPD にはその転写系が、RNA polymerase II によるものかを実験的に示されたデータが載っており、これから今回最後に特徴づけた配列が RNA polymerase II 系かどうかの同定に使用している。

4　結果

今回用いたデータセットと抽出出来たコンセンサスパターンの数を示した(Table 1)。この表から抽出できたパターンの数はデータの数には依存していないことがわかる。また、データセットとしてEMBL の区分を用いたため、invertebrate のような非常に広い生物種を含む様なものでは抽出できたコンセンサスパターンの数は少なくなる。その反対に prokaryote のデータでは、データ数が少ないにもかかわらず、ほぼ同様の体制の生物種から構成されているためか、抽出されたコンセンサスパターンの数はデータ数の多かったものと変わらなかった。

最終的に統合的に転写上流配列を特徴づけている例を示す(fig. 2)。ここでは各遺伝子の上流域を、コンセンサスパターン検索では 1 塩基置換を許容したパターンマッチングを、また TFD による検索では完全マッチだが、反対側のストランドも同時に検索している。この例のように普通よく知られているTATA box 様の配列であっても実験で全ての類似パターンが同定されているわけではない。そのため、実験的データのみによるパターンマッチングでは重要な配列を落しうる可能性が高い。しかし、これを統計的なデータを加えることである程度補足しうることがわかった。

5　生物学的意義と考察

今回の特徴づけではかなりの程度転写上流域を特徴づけられはしたが、統計的手法には位置固定性のないパターン抽出を折り込んでいない。今後は非位置固定性のパターンも特徴づけに加味していくべきであろう。

現在これらの特徴づけをした配列に対して機能面からの分類を検討中である。その遺伝子の発現する時期が、例えば housekeeping gene のように常時であるもの、一過性のもの、組織特異的であるものなどによって予測が出来れば研究上大いに役立つものと思われる。

本研究は科学研究費重点領域研究「ゲノム情報」の助成を受けている。また京都大学化学研究所スーパーコンピューターラボラトリーより計算機の CPU time の提供を受けた。

参考文献

1) D.J.Galas, M.Eggert & M.S.Waterman.:J.Mol.Biol.,186,117-128(1985).
2) D.Ghosh.:Nucleic Acids Res.,18,1749-1756(1990).
3) P.Bucher & E.N.Trinov.:Nucleic Acids Res.,14,10009-10026(1986).

```
[ -38 -32 ] CCGGCC 224
[ -38 -38 ] GGGGAG 27
[ -37 -37 ] GGGGCG 38
[ -37 -37 ] GGCTCT 27
[ -37 -37 ] TATAAA 25
[ -37 -37 ] CTCCTC 25
[ -36 -29 ] GGGCGG 270
[ -36 -36 ] CTATAA 32
[ -36 -36 ] CTCTCT 24
[ -36 -36 ] GGCTCC 24
[ -36 -36 ] AGTATA 24
[ -35 -24 ] TATAAA 1341
[ -35 -28 ] GCATAA 219
[ -34 -34 ] GACATA 24
[ -28 -28 ] CGCGCG 24
[ -27 -25 ] AAGGCC 100
[ -27 -24 ] AGGCAG 130
[ -27 -27 ] AAGGGG 29
[ -27 -27 ] ACATAA 24
[ -26 -20 ] CGCGCG 196
[ -26 -26 ] GGCTGC 24
```

Fig. 1. The consensus pattern index. Locational information from the transcriptional intiation site, patterns and frequencies are indicated

Table 1. Data sets and extracted consensus patterns.

Data set	*region*	*number*	*consensus*
emblpri.dat	*-100 ~ +9*	*233*	*254*
emblrod.dat	*-100 ~ +9*	*239*	*259*
emblinv.dat	*-100 ~ +9*	*112*	*121*
emblpln.dat	*-100 ~ +9*	*110*	*282*
emblfun.dat	*-100 ~ +9*	*88*	*244*
emblpro.dat	*-100 ~ +9*	*75*	*212*

```
Entry:                  AGHBD + 235
Polymerase type:        polII
Product:                Spider monkey (A.geoffroyi) delta-globin gene,
                        complete cds.
Retrieved Pattern:      (3/6)AACCAAT(4/7)CTTATCTT(5/7)CCTGC(1/1)CAGAGG
                        (1/1)GCAGG(0/1)AGAACAGGACC(1/1)GCATAAAAGGCAGGGCAGGG
                        (0/1)TAACTGTT(0/1)CTTTGACTT(2/4)
```

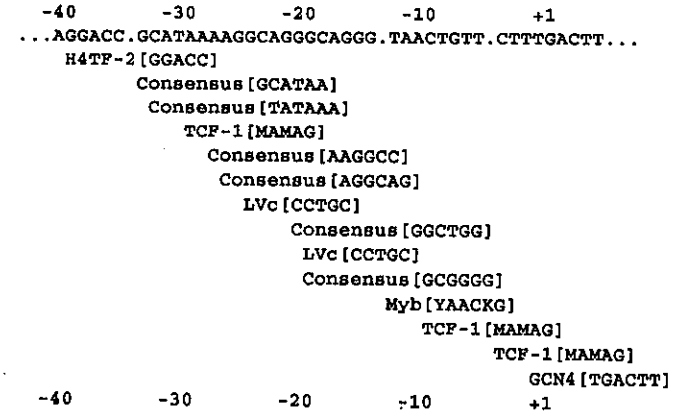

Fig. 2. An example of integrated retrieval with consensus patterns and experimental data.