Computer assisted identification and classification of streptomycese promoters

William R. Bourn* and Brendan Babb

Departments of Microbiology and Medical Microbiology, University of Cape Town, Rondebosch 7700, Cape Town, South Africa

Received June 2, 1995; Revised and Accepted August 17, 1995

ABSTRACT

Short sequences that were over represented in a database of Streptomyces promoter region sequences were identified. These sequences and others that were selected on the basis of the characteristics of known promoters, were tested to determine if they were found predominantly at particular distances from the transcription start site. In several cases obvious clusters were recorded. This has allowed the objective identification of potential promoter core sequences. In some cases these may define novel promoter classes. 150 Streptomyces promoters have been listed and grouped on this basis. A new and extended consensus sequence for the Streptomyces E.coli σ^70-like promoters was determined. It showed differences from that of E.coli, both in sequence and in the spacing between the −35 and −10 regions.

INTRODUCTION

The genetic regulation of the streptomycetes is of interest primarily because this genus produces the majority of the world’s antibiotics. This antibiotic production is intimately linked with the complex process of development (1,2).

Most bacterial genes are regulated at the level of transcription, often in part by the controlled expression of different RNA polymerase σ factors, which recognize different classes of promoter. This enables the tight regulation of multiple promoters of a particular class and is used for the regulation of genes as diverse as those involved in nitrogen fixation, flagella production, heat shock response and sporulation (3–6). Typically, bacteria have a major σ factor, responsible for the recognition of the majority of the promoters, while other minor σ factors have limited and specialized functions.

A list of 139 Streptomyces promoters has been previously compiled (7). A number of the genes listed are transcribed from multiple promoters and/or have overlapping divergent promoters. Several promoters are found within protein coding sequences. Approximately 21% of the promoters were designated as E.coli-like Streptomyces promoters, on the basis of an undefined similarity to the promoters recognized by the E.coli σ^70 (7,8). No other separate classes of promoters were obvious among those remaining.

Different experimental approaches have proved, or implied, the existence of multiple Streptomyces σ factors (9). The first demonstration utilized S.coelicolor proteins and Bacillus subtilis promoters (10). Two σ factors, σ^35 and σ^49, were isolated and in vitro these directly RNA polymerase recognition of the veg and ctc promoters respectively.

The existence of another S.coelicolor σ factor was demonstrated with the four heterogeneous promoters, dagAp1−4 (11). Three separate transcribing activities for each of dagAp2, dagAp3 and dagAp4 were distinguishable. In vitro transcription proved that dagAp2 was recognized by the novel σ^28 (12). The gene encoding σ^28 (sigE) has been sequenced and appears to encode a σ factor of a class thought to regulate extracytoplasmic functions (13). Transcription activity co-purification studies utilizing the veg, ctc and dagA promoters indicated that dagAp3 was recognized by σ^49 while dagAp4 was transcribed using σ^35 (9).

Studies on the gal operon have implied that at least two other σ factors, in addition to σ^35 and σ^49, exist in S.coelicolor. Neither of the two gal promoters was recognized by σ^35 or σ^49 in vitro and the transcribing activities of the two promoters could be separated (14,15).

To identify the principal S.coelicolor σ factor gene, an oligonucleotide probe consisting of sequence conserved only in the major σ factor genes was used in Southern blots (16). Four homologous regions were identified and subsequently sequenced (17,18). Each of the four loci, termed hrdA−D, encodes a homolog of the major σ factor class. Important residues are conserved in the 2.4 and 4.2 regions of the hrd genes, indicating that the σ^Hrd proteins recognize subtly different promoter sequences (9,19). Transcription of Streptomyces hrd homologues appears to be dependent upon the growth medium or developmental stage (19,20). Mutational studies indicated that hrdA, hrdC and hrdD are non-essential for growth, while hrdB is crucial and encodes the equivalent of σ^10 of E.coli (19,21). Only hrdB has been shown to produce a viable σ sigma factor, of 66 kDa. This can direct transcription from the veg and dagAp4 promoters (22). It is unclear whether the 35 kDa σ^35 is a different σ factor or a breakdown product of σ^HrdB. There is no evidence that the other hrd genes are translated. However the transcribing activities for the veg promoter and the S.lividans XP55 promoter have been separated, although these promoters have a high degree of
The fraction of the total number of nucleotides in the database by the expected number of occurrences when assuming each of the different octanucleotides was generated. For each (\%A, \%G, and \%T respectively). The frequency of each \%C which consisted of each of A, G, C and T residues was calculated frequencies of monomers to heptamers were determined from the normalized form, by dividing the actual number of occurrences octanucleotides was catalogued. An eight base window was used for transcription start site. Potential promoter core sequences were therefore tested to determine if they were found predominantly at set distances from the transcription start sites.

### MATERIALS AND METHODS

**Creation of catalogues of oligonucleotide frequencies for *Streptomyces* promoters**

The following analyses were conducted using computer programs written by ourselves. The programs were written in Borland C++ and ran with MS DOS.

Using *Streptomyces* promoter sequences, the frequencies of occurrence of each of the 65 536 different possible overlapping octanucleotides was catalogued. An eight base window was moved, one base at a time, down the sequences and at each position the octanucleotide sequence was determined. A tally of each of the different octanucleotides was generated. For each possible octamer the frequency of occurrence (F) was calculated in normalized form, by dividing the actual number of occurrences by the expected number of occurrences when assuming completely random sequence. The number of occurrences and frequencies of monomers to heptamers were determined from the octamer data.

\[
F = \frac{\text{number of occurrences}}{\text{total number of nucleotides in database}} \times \text{oligonucleotide length}
\]

The fraction of the total number of nucleotides in the database which consisted of each of A, G, C and T residues was calculated (\%A, \%G, \%C and \%T respectively). The frequency of each oligomer, corrected for the nucleotide ratio (NCF) was then calculated using the number of times each base was found in the oligonucleotide (XA, XG, XC and XT). An NCF value of 1 therefore signifies that the oligomer occurs at the expected frequency, taking into account the nucleotide bias. Higher or lower values indicate that the oligomer is either over or under represented.

\[
NCF = \frac{F(XA)}{(%A)} \times \frac{F(XG)}{(%G)} \times \frac{F(XC)}{(%C)} \times \frac{F(XT)}{(%T)}
\]

Two different oligonucleotide frequency databases, termed Spro#1 and Spro#2, were created, using *Streptomyces* promoters that have been identified by transcription mapping (Table 1). Each promoter sequence was analyzed individually, without concatenation. Where there exists a number of different overlapping promoters for a given gene for example *amy*-p1, *amy*-p2 and *amy*-p3 (Table 1, promoters number 7, 140 and 132), these were entered in the database as a single contiguous sequence. Only one strand of the DNA was entered. Consequently, in the case of overlapping divergent promoters one of them was entered as reverse complementary sequence. Only sequences between positions –53 and +1, as defined in Table 1, were used. No known protein coding sequence was entered. The Spro#1 database consisted of all promoters listed in Table 1 which, as far as we could establish at the time of the creation of the database, fulfilled these criteria. The promoters included in the Spro#2 database fulfilled both these and other criteria (see Results and Discussion). Spro#1 and Spro#2 consisted of 3959 and 1567 bases, entered as 106 and 37 different contiguous DNA sequences respectively. These figures do not include the last seven bases of each sequence entered and are equivalent to the total number of octamers counted.

**Promoter region sequence bias cluster analysis**

A particular oligonucleotide was selected for testing if it had a high NCF value and was thus greatly over represented in the database being examined. The minimum NCF value considered in each case was set such that manageble numbers of over represented oligomers (up to 70) of each type were identified. Other sequences were selected for testing using different criteria (see Results and Discussion). In the search for the former, gaps (positions that could represent any base) were inserted in the oligomer sequence. The length of oligomer examined and the gap pattern used are shown in Table 2. The lengths and patterns used were limited. This was because degenerate sequences that are effectively dimers will contain limited information; sequences which are effectively pentamers or larger generate unmanageable amounts of output-data-and-prone to NCF value fluctuation due to the limited size of the database.

A number of *Streptomyces* promoters were aligned at the point of transcription initiation. Only sequences between and including positions –47 and +1 of the promoters were utilized. Where there were multiple initiation points the nearest to the promoter was considered as position +1. Two groups of promoters were used on separate occasions (see Results and Discussion). Group 1 consisted of all the *Streptomyces* promoters listed in Table 1 except for *endoH*-p, *orfRP*-p, *vlfF*-p, *blaF*-p, *octD*-p4, *octD*-p5,6, *vph*-p1, *vph*-p2 and *kmr*-p (promoter numbers 5, 20, 59, 89, 120, 134, 135, 141 and 142); these were not included as their sequences did not extend as far as position –47 or the data was not available at the time of cluster analysis. Group 2 consisted of the
promoters listed in Table 1 numbered from 90 to 150 except for oclD-p1, oclD-p5, 6, vph-p1, vph-p2 and kmr-p.

The oligonucleotides selected were tested against either the Group 1 or 2 promoters. Each occasion they occurred within any promoter region was recorded. Also recorded was the position of the first base of the oligonucleotide in relation to the transcription start site of the promoter in which it was found. The number of times that an oligomer under consideration occurred at a particular distance (measured in numbers of bases) from the transcription start sites of the promoters was plotted as a bar graph. Putative promoter recognition sequences were aligned and the total number of times each base occurred in each position was calculated.

RESULTS AND DISCUSSION
Promoter region sequence bias cluster analysis

Promoter region sequence bias cluster analysis was conducted on a number of *Streptomyces* promoters. Initially the selection of the
Table 2. Type and number of oligomer used for promoter region cluster analysis

<table>
<thead>
<tr>
<th>Type</th>
<th>Using Database SPRO#1</th>
<th>Using Database SPRO#2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIMERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNN</td>
<td>U</td>
<td>T 1.6 30</td>
</tr>
<tr>
<td>TETRAMERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNNN</td>
<td>T 2.0 6 17.3 33</td>
<td></td>
</tr>
<tr>
<td>NNN-N</td>
<td>U 1.2 16</td>
<td></td>
</tr>
<tr>
<td>NN-N</td>
<td>U 1.2 15</td>
<td></td>
</tr>
<tr>
<td>PENTAMERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NNNNN</td>
<td>T 2.8 13 3.5 32</td>
<td></td>
</tr>
<tr>
<td>N-NNN</td>
<td>U 1.5 36</td>
<td></td>
</tr>
<tr>
<td>N-N-N</td>
<td>U 1.6 33</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N</td>
<td>U 1.5 45</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N</td>
<td>U 1.2 13 1.2 9 1.2 11</td>
<td></td>
</tr>
<tr>
<td>HEXAMERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N-N-NNN</td>
<td>U 1.5 6 1.5 37</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N</td>
<td>T 1.5 18 1.5 39</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N</td>
<td>T 1.7 10 1.6 39</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N</td>
<td>T 1.5 16 1.4 51</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N</td>
<td>T 1.5 15 1.4 42</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N-N</td>
<td>T 1.3 15 1.3 43</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N-N-N</td>
<td>T 1.3 5 1.1 21</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N-N-N-N</td>
<td>T 1.2 9 1.1 22</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N-N-N-N-N</td>
<td>T 1.2 10 1.1 24</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N-N-N-N-N-N</td>
<td>T 1.2 5 1.1 10</td>
<td></td>
</tr>
<tr>
<td>N-N-N-N-N-N-N-N-N-N-N-N-N</td>
<td>T 1.0 6 1.0 10</td>
<td></td>
</tr>
</tbody>
</table>

aSymbols: N, represents in turn each of A, G, C or T specifically; −, represents any base and is recorded as a space.

bSymbols: T, NCF value of oligomers was examined; U, NCF of oligomers was not examined.

The minimum NCF value required for an oligomer to be tested by cluster analysis.

The number of oligomers that had required NCF value and were tested by cluster analysis.

oligomers, of up to six bases, to be used for the test was made using the Spro#1 database and NCF data. However the database was too small to support the use of hexamers without any gaps in the sequence. To overcome this, the oligomers used were such that not every position within their length was assigned as either A, G, C or T. The databases used, types and numbers of oligomers examined and the minimum NCF value required for them to be considered for testing are listed in Table 2.

Cluster analysis was conducted using the Group 1 promoters unless otherwise specified. Selected results of the cluster analysis are shown in Figure 1. Many of the oligomers gave clear or highly suggestive results. These sequences have been termed potential promoter core sequence(s) (PPCS) and grouped into Classes A–H. Each Class consisted of either a single PPCS or a number of PPCS which were selected for testing by the same logical process and were similar in sequence.

The most striking result was for the sequence TANNNT (PPCS Class A; Fig. 1, A1). Of the 67 times this sequence occurred, 50 times (74.6%) it was found in the region −14 to −10. The clustering of this sequence could be due to the presence of E.coli σ70-type promoters. However, inspection of the results from cluster analysis using the sequences TANNGT and TANNAT indicate that the preferred −10 region varies from the E.coli consensus (data not shown).

A consensus calculation was performed using all the PPCS Class A sequences that fall between positions −14 and −10 (Table 3A). There are 52 such promoters, including orfRP-p and rylp-p, which were not used in the cluster analysis because of their limited length. It is difficult to determine a consensus sequence for an organism, such as Streptomyces, that has a strong G/C bias because this will tend to obscure the base preference due to the function of the DNA sequence under consideration. However, it would be expected that for positions within the consensus that have no function, G and C residues would be equally represented. A and T residues should likewise be equally represented. Wherever this was not the case it has been assumed that this is due to a base preference required for the protein binding activity of the DNA. The resulting consensus, TA(G/C)(G/C/A)(G/T)T, differs considerably from that of E.coli. Most notable is the fact that T is the least common base at position −11 while A is least common at position −9, in direct contrast to the situation in E.coli (inverted commas are used to indicate that the ‘position’ is as defined for Table 3).

The PPCS Class B is less obviously clustered than Class A (Fig. 1, B1). In this case the sequence CANNOT occurs in the −15 to −12 region 10 times out of a total of 16 (62.5%). It is the absence of this sequence in other regions that is highly suggestive, because it is likely that promoter-like regions will be avoided at positions where they have no function.

It could be argued that PPCS Class B merely represents a subset of promoters that are transcribed by the same RNA polymerase that recognizes Class A. However, it is reasonable to expect that were that the case, a sequence that retained features of Class B, but was closer to the consensus sequence of Class A, would be found clustered in the −10 region. CANNTG is such a sequence, yet in this case (and others, data not shown) no such clustering is
observed (Fig. 1, B2). Thus the sequence CANNAT may represent the core components of a novel -10 region.

The PPCS Classes A and B appear to be associated with a third class, C, in the -35 region (Fig 1, C1). This sequence, TTGAC, is found over 5-fold more often than predicted (database Spro#1, NCF value 5.43). 14 out of 20 (70%) sequences of this class start in the -38 to -33 region. Furthermore, the sequence is exactly that of the E.coli major σ factor -35 recognition site consensus sequence, except one base shorter (Table 3B). This strongly indicates that the cluster represents the -35 recognition site of Streptomyces E.coli-like promoters. It should also be noted that when the PPCS was found outside the cluster, this was due to the phenomenon of overlapping multiple promoters (as was often the case with other PPCS classes, data not shown). Thus, the occurrence of the sequence TTGAC at positions -10 and -18 is as a result of the same sequence being found at different positions in the amy-p3 and amy-p2 promoters as well as in the amy-p1 promoter at position -36 (Table 1, promoter numbers 132, 140 and 7). The same is true of the bpaA-p1 and bpaA-p2 promoters at positions -33 and -26 (Table 1, promoter numbers 5 and 103).

Several attempts were made to identify more of these sites using limited subsets of the sequence TTGAC, however this was unsuccessful (data not shown). Although clusters were obvious it was impossible to distinguish in which instances the test sequence was likely to be functionally significant. This was due to the frequent occurrence of the same sequence in regions other than the -35 region. An example of this is shown for the sequence TTG (Fig. 1, C2).

A consensus sequence for the 15 PPCS Class C sequences found between positions -38 and -33 (including that of endoH-p, which was not included in the cluster analysis) was generated (Table 3B). Although there are too few of these sequences to be confident that small deviations are meaningful, it should be noted that there appears to be highly conserved A, G, C and T residues at positions -41, -38, -37 and -30 respectively. This consensus is markedly different from that of E.coli in several...
features. Most importantly, the A residue at position ‘-31’, which is highly conserved in E.coli, does not appear to be exceptionally common in Streptomyces.

Another class of hexamer (CNGNNA) was seen to be clustered between bases -18 and -15 (Fig. 1, D1). The sequence was often associated with, and overlapped with PPCS Classes A and B. This suggests a functional role for the sequence CNGNNA because it has only one base in common with those two other classes. In addition, the consensus sequence for the PPCS Class A (Table 3A) shows that C and G are predominant in positions ‘-17’ and ‘-15’ respectively. This observation prompted a search for other oligomers clustered upstream of the -10 region, selected on the basis of their similarity to the extended PPCS Class A consensus sequence. Several were found and an example is shown for the sequence TGNNNA (Fig. 1, D2). It is interesting to note that in this case there is a second cluster in the -35 region which represents a subset of PPCS Class C.

Thus the cluster analyses and the consensus sequence data together suggest that, as in E.coli (28), an extended consensus sequence occurs in the -10 region in Streptomyces. Further cluster analyses were performed using sequences that were based on the PPCS Class A consensus in the ‘-18’ to ‘-8’ region. Again, some degree of clustering was recorded for a number of oligomers but only four were free from a high background. These have been termed PPCS Class D, and include the sequences TNNNNANNNT, TNGNANNNT, CNGNNANNNT and TGNNANNNT. Both the least and most obvious clusters of this type are shown (Fig. 1, D3 and D4).

It could be argued that the clustering of PPCS Class D simply reflects the fact that the sequence ANNNT itself occurs as a cluster (data not shown). Two lines of evidence suggest that this is not the case. Firstly, it would then follow that the clustering of the sequences TGNNA and CNGNNA arises due to a preponderance of A residues in the -12 region. This would mean that most sequences that contain an A would be clustered in this region, which is not the case (data not shown). Secondly, if the ANNNT motif of the Class D PPCS is retained intact, while the remaining nucleotides are assigned new positions, no peak is observable after cluster analysis. As an example of this the results for the sequences TGNANNNT and GTNANNNT can be compared (Fig. 1, D4 and D5).

Of the 150 promoters tested, 66 (Table 1, promoter numbers 85–150) did not fail into any of the classes that were defined using the Spro#1 database to select over represented sequences. There are several possible reasons for this. First, it is possible that there are other classes of promoter which were not identified as they are rare and so do not appear as an obvious peak in cluster analysis. Such a problem would be exacerbated by the presence of large numbers of promoters of the major class(es) in the original sequence data entered. Alternatively, it is possible that the oligomers that define more common promoter recognition core sequences are not over represented in the database and were not tested because of this. This could occur if the over representation due to the oligomers appearance in the protein binding regions is counterbalanced by (not unexpected) under representation in other parts of the promoter. It is also possible that other PPCS were not identified because they are G/C rich and because the frequency of occurrence has been corrected for nucleotide bias. Any G/C rich oligomer would have to be numerically more highly represented than an A/T rich sequence in order to attain a similar NCF value. Finally it is possible that alternative PPCS classes were not identified because only hexamers were extensively tested.

Alternatively, many of the promoters that have not been classified on the basis of cluster analysis are probably members of the same promoter class(es) as those that have been tentatively defined above. In these cases, the -10 and -35 regions might vary too much from the consensus to be identified by the method used here.

In order to address these problems a number of approaches were taken. First the biochemical data concerning certain promoters were considered in regard to the cluster analysis results. There are a limited number of Streptomyces promoters which are known to be expressed via different transcription factors. These include the promoters for the dagA and gal operons, the XP55 promoter and the σwhlG dependent promoters PTH4 and PTH720 (promoter numbers 21, 33, 76, 88, 43, 124, 1, 107 and 109). As only dagA-p2, gal-p2 and the σwhlG dependent promoters do not contain sequences that can be classified as any of the PPCS Classes A to D, it is possible that these are representatives of minor classes.

The consensus sequence of the σwhlG dependent promoters is thought to be similar to that of the motility related promoters of other bacteria or the sequence TAAA(N15)GCCGATA(A/T) (29). In the case of the dagA-p2 –35 region recognition site there is evidence that the consensus may be CCGGAACCTT (13). Cluster analysis of all possible subsets of these sequences (triplets to octamers, with and without appropriate blank spaces) was carried out. No clusters were observed using any of the σwhlG target-type sequences. This is not surprising as there has been no widespread effort to isolate and map this type of promoter. Furthermore, the majority of the transcription mapping of the promoters used in this study was performed using logarithmically growing mycelia from liquid medium, in which σwhlG is probably not expressed. These two factors would ensure that σwhlG dependent promoters are poorly represented among the promoters examined here.

In the case of the dagA-p2 –35 region-type sequences, only one, GNAAAC, gave any indication of clustering, albeit very weakly (Fig. 1, E1). This sequence has been termed PPCS Class E and its significance is dubious. It is interesting to note, however, that in three cases (hrdD-p1, whlB-p1 and dagA-p2, promoter numbers 86, 87 and 88) there is a correctly spaced TC dimer in the -10 region, as the putative recognition sequence for the σ28 promoter class suggests there should be (13). This observation was used to define a further class of PPCS (see below).

To address the possibility that the presence of many of the major class(es) of promoters in the Spro#1 database was hindering the detection of minor classes, the Spro#2 database was created. This consisted only of promoters that did not contain appropriately positioned PPCS Classes A–E (Table 1). As with the Spro#1 database, highly over represented sequences in the Spro#2 database were identified (Table 2). These were tested against the Group 2 promoters, which do not include among their number any promoters containing appropriately positioned PPCS Classes A–E, using triplets to hexamers.

Only one sequence, GNTTNC, gave results that might be interpreted, with caution, as a cluster (Fig. 1, F1). It is interesting to note, however, that when this sequence (Class F) was tested using the Group 1 promoters, the cluster pattern was almost identical (Fig. 1, F2), indicating that the sequence is very rarely
found among the promoters that contain appropriately positioned PPCS Classes A–E.

Further tests were conducted to attempt to isolate promoter-like sequences that resembled the previously identified PPCS classes poorly, but which might still be functional. It was argued that while the −10 and −35 regions alone could not be recognized, together there could be enough sequence information for their identification. As the PPCS Class A and B sequences appear to be associated with those of Class C, subsets of these classes, with a variety of spaces between them were tested.

To first establish the limits of the correct spacing, the sequences TTG and ANNNT (which are highly conserved in E.coli) were used, with the spacing between them set at 20–24 bases. Only in the cases of a 21 and 22 base separation were clusters recorded, indicating that this was the correct spacing. It should be noted that this represents a spacing of 18 and 19 bases between the TTGAC and (T/C)ANNNT and is similar to the spacing found for the E.coli major class of promoter. However, the 17 base spacing found in E.coli was not observed (Fig. 1G1–G3).

With the spacing between the −10 and −35 regions set at either of these two distances, all possible combinations of subsets of TTGAC and (T/C)ANNNT were tested. Thirteen sequences, TTG(N20)TNNNT, TGA(N19/20)TNNNT, TGA(N20)ANNNT, GAC(N18/19)TNNNT, GAC(N19/20)ANNNT, TNGA(N20)ANNNT and TNGA(N19)TNNNT showed some clustering (Class G). The cases where this is the most and least obvious are shown (Fig. 1, G4 and G5).

A similar approach was adopted to test for a −10/−35 association within the PPCS Class E sequences. As noted above, inspection revealed that in some cases there appeared to be an appropriately positioned TC dimer in the −10 region of promoters that fall within this class. For this reason cluster analysis was conducted on the sequences GNAAC(N19)T and GNAAC(N20)T. While similar sequences with lesser or greater spacing between the −10 and −35 regions were tested and did not show clustering (data not shown), when the spacing was either 19 or 20 bases, clusters could be observed (Fig. 1H1 and H2). These two sequences have been termed PPCS Class H. The clustering that is observed strongly indicates that these sequences represent a distinct class of promoters, and gives support to the contention that the indistinct clustering of PPCS Class E sequences is not coincidental.

The promoters that fall into the various classes are shown in Table 1. 59% of the promoters can be described as members of the Classes A, B, C, D and G, and it seems that these represent E.coli-like promoters. Within this group are found the XP55-p, dagA-p4, dagA-p1, dagA-p3 and gal-p1 promoters (promoters number 1, 24, 33, 76 and 43). Also in this group are the B.subtilis veg and ctc promoters (Table 1). It is interesting to note that although the promoters are so similar, there are at least three distinct transcribing activities that recognize different members of this group. Perhaps these consist of (at least in part) the different hrd homologues. Other E.coli-like promoters of note include whiB-p2 and hrdB-p (promoters number 9 and 82; see below). In the case of whiB-p2 the spacing between the −35 and −10 regions is very unusual (21 bp).

Promoters in the Classes E and H might be of the σ28 class. This is particularly relevant because these include the promoters of a number of fundamentally important genes including hrdB and hrdD (promoters number 82 and 86). In the case of hrdB-p, Class B and H sequences overlap, suggesting that hrdB could be transcribed from either an E.coli-like or σ28 type promoter, or indeed from both. Also of note is the fact that whiB-p1 (promoter number 87) appears to be a σ28 type promoter.

The blaF-p promoter (number 89) contains PPCS Class E and H sequences, although these are not within the cluster but three bases removed from it. However extensive mutation analysis of this promoter has been performed (27) and comparison of these results with those obtained here strongly suggest that blaF-p is a σ28 type promoter. Both point mutation and deletion analysis are compatible with the results reported here.

The gpl-p2 promoter (number 124) is an interesting case as it contains both the PPCS Class E and H sequences, but in the wrong position, at −45. Inspection of this promoter reveals that it matches the putative consensus sequence for the σ28 (13) in 10 of the 11 positions, with the correct spacing between the −10 and −35 regions. It seems unlikely that such a sequence could occur by chance (especially as it is relatively A/T rich and the database is so small) and that it should have no function.

There are no well studied promoters that fall in the PPCS Classes B and F. It is also difficult to determine the significance, if any of these classes.

It is possible that the sequence bias recorded here is as a result of recent gene duplication. However none of the reports from which the sequence data were obtained describe any extensive homology with other promoters. It follows therefore that any gene duplications included in the data here were either limited or not recent.

Although the study of genetic regulation in Streptomyces is in its infancy, it is already clear that there are a multiplicity of promoter types. This will certainly have bearing not only on the general metabolism of the bacteria, but also on their development and antibiotic production. To date there have been no formal studies, using sequence databases, on what constitutes a Strepto-myces promoter. While it has been possible to recognize those promoters that bear some similarity to the E.coli major class of promoters, the criteria for this are not defined and there has been no way to estimate the probability that such assignments are correct. This cluster analysis was aimed, in part, at addressing this and determining where the differences between the major promoter classes of Streptomyces and E.coli lay. The approach is clearly subjective to a degree, particularly with regard to defining what constitutes a cluster in cases where it is not immediately obvious. However, further statistical analysis must ultimately rely on pre-set conditions, the selection of which is also subjective. The results reported here are therefore not intended to be a definitive classification but should rather serve to direct the researchers attention to potential promoter core sequences.

The use of cluster analysis to identify promoters has great potential, particularly as the list of sequenced and mapped Streptomyces promoters is growing rapidly. This will also have bearing on the genetics of other actinomycetes, and indeed other organisms.

ACKNOWLEDGEMENTS

We are indebted to J. A. Thomson and R. Kirby for advice and critical contributions to this work.

REFERENCES


23 Westpheling, J. and Bawrner, M. Reported in reference 9.