Nucleic Acids Research, 1993, Vol. 21, No. 3 733–739

Quantitative sequence-activity models (QSAM)—tools for sequence design

Jörgen Jonsson, Torbjörn Norberg[1], Lena Carlsson[1], Claes Gustafsson[1] and Svante Wold
Research Group for Chemometrics, Department of Organic Chemistry and [1]Department of Microbiology, University of Umeå, S-901 87, Umeå, Sweden

Received May 18, 1992; Revised and Accepted December 29, 1992

ABSTRACT

Models have been developed that allow the biological activity of a DNA segment to be altered in a desired direction. Partial least squares projections to latent structures (PLS) was used to establish a quantitative model between a numerical description of 68 bp fragments of 25 *E.coli* promoters and their corresponding quantitative measure of *in vivo* strength. This quantitative sequence-activity model (QSAM) was used to generate two 68 bp fragments predicted to be more potent promoters than any of those on which the model originally was based. The optimized structures were experimentally verified to be strong promoters *in vivo*.

INTRODUCTION

We are here concerned with the relation between the composition of a DNA sequence and its associated biological activity. The analysis of sequence data has traditionally been concentrated on qualitative pattern recognition (i.e. classification). This involves, for example, models that are based on the observed similarity (i.e. homology) between sequences (1–5). Homology based models have also been used in attempts to model the magnitude of functional properties of sequences (6). Such models have, however, been criticized for being of limited predictive value (7, 8). Here we aim to demonstrate that sequence data may carry two complementary pieces of information. The first part is the homology, i.e. information related to absence of variation. The second, less well recognized information is that conveyed by systematic variation. For quantitative correlations in a class of related sequences, the information based on systematic variance must also be extracted and utilized.

The reason that potential co-variance structures are usually not considered is that the potential of multivariate methods like e.g principal components analysis (PCA), partial least squares projections to latent structures (PLS) and neural networks (NN), for sequence modelling purposes has not been recognized until recently (9–11). NN have successfully been used to classify digitized DNA sequences, see e.g. (12 –14). The applicability of PLS to quantitative sequence-activity modelling (QSAM) has been addressed by us in earlier papers (10, 14–16). It is thus important to distinguish between QSAM and classical sequence pattern recognition modelling (see refs. above). In the present context the QSAM is developed within a class of functionally related sequences. The objective is to delineate the relationship between the sequences and the magnitude of their corresponding biological activity.

The aim of this paper is, however, not primarily to design a strong bacterial promoter. The objective is rather to outline a general strategy whereby the relationship between the composition of a bio-polymer and its biological activity may be quantitatively described and subsequently utilized. It should be noted that models of the category presented here are local linearizations of the more complex functions underlying the biological phenomena observed. Consequently, QSAMs are of local validity, i.e. interpretations and predictions relate to the experimental conditions used to characterize the set of sequences upon which the models are based.

Parametrization of DNA sequence

A numerical representation of the sequence is a prerequisite for quantitative modelling. One possibility is to use a qualitative (discrete) parametrization of the monomers. This corresponds to the use of indicator variables that unequivocally and symmetrically state the identities of the relevant monomers. This implies that a minimum of three descriptor variables/base must be used in order to obtain an informationally efficient representation for DNA (see refs. 10 and 12). Another alternative is given by quantitative monomer parameters, i.e. continuous variables, such as principal properties (PPs) derived from measured physico-chemical data collected for monomers of interest (15, 16). These two kinds of descriptors have different advantages. The qualitative indicator variables are conceptually simpler, easier to derive and more readily interpreted. Properly derived quantitative descriptors may, however, enable an interpretation in terms of which physico-chemical factors that are important for the biological response and how they combine etc. In this example we have utilized four discrete indicator variables to represent the bases of DNA (A=1000, C=0100, G=0010 and T=0001). The reason for this selection is that the resulting model parameters are the least complicated to interpret.

Promoter sequence data

Prokaryotic transcriptional promoters are specific DNA sequences that are recognized by the σ-unit of the RNA polymerase holoenzyme (RNAP). The assembled enzyme complex

subsequently initiates and carries out the transcription of mRNA from the DNA template. There are many examples of sequences known to be functional *E.coli* promoters. Some originate from the bacterium itself, others from infecting phages. Few of these promoters have been consistently characterized with regard to their *in vivo* promoter strength. However, a system that allows this efficiency parameter to be determined relative to an internal standard has been developed and used to characterize promoters by Bujard and coworkers (5, 6, 15–17). In this assay the strength of the test promoter (in front of the hydrofolate reductase gene, *dhfr*) is expressed relative to that of the promoter for β-lactamase (P_{bla}) which is present on the same plasmid. Monitoring of the mRNA expressed from the promoter under study in relation to the standard, permits the relative promoter efficiency to be determined unbiased by translational effects or gene dosage. These data were considered to be comparatively well suited for QSAM development for two reasons; **a)** this material comprises a relatively large set of structures that are multipositionally altered, and thereby informationally better suited than similar sets generated using saturation mutagenesis, and **b)** the better comparability for additional structures resulting from an experimental protocol comprising both an internal standard and external references.

Analysis of sequence data is normally based on the assumption that certain positions in the sequence in some way interact with a target molecule. This, in turn, corresponds to the requirement that the sequences to be analyzed are of similar length and that they are properly aligned. Modifications of these requirements may, for example, be made by dividing the sequence into subsequences around given points of reference. However, structural descriptions that are alignment independent may also be accomplished, e.g. according to the principles outlined by van Heel (9). In this paper the principles of multivariate DNA QSAM are illustrated using the traditional alignment dependent representation of sequence data. The present models are hence based on promoters having similar distances between the positions −35, −10 and +1 this, consequently, makes each position of the 68 mer to be more directly comparable. From references (5, 6, 15–17), the 68 bp fragment (−49 to +19) relative to the start of transcription was compiled. All promoters having a 17 bp spacer between −35 and −10 region and a 7 bp spacer between −10 and +1 region were considered. This subset comprising 25 promoters was found to comprise three major categories; 1) $P_{D/E20}$, P_{G25}, P_{J5} and P_{N25} from phage T5, 2) P_L from phage lambda and, 3) P_{con} an artificial construct originally synthesized by Dobrynin et al. (18). The structures of promoters

Table 1.

Promoter	Strength (log P_{bla}-units)		Promoter sequences (−49 −40 −30 −20 −10 +1 +10 +19)
1 $P_{D/E20}$	1.748	1	ACTGCAAAAATAGTTTGACACCCTAGCCGATAGGCTTTAAGATGTACCCAGTTCGATGAGAGCGATAA
2 P_{G25}	1.278	2	GAAAAATAAAATTCTTGATAAAATTTTCCAATACTATTATAATATTGTTATTAAAGAGGAGAAATTAA
3 P_{J5}	0.954	3	TATAAAAACCGTTATTGACACAGGTGGAAATTTAGAATATACTGTTAGTAAACCTAATGGATCGACCT
4 P_{N25}	1.477	4	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT
5 $P_{N25/O3}$	0.895	5	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAATTGTGAGCGGATAACAA
6 $P_{N25/O4}$	1.246	6	CATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAATAGATTCATAAATTTGAGAGAGGAGT
7 $P_{N25/O5}$	1.173	7	GGATAACAATTTAGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT
8 $P_{N25/pex}$	1.176	8	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAAGGGTCGAGAAGAGT
9 $P_{N25/aDSR}$	0.431	9	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATCCGGAATCCTCTTCCCG
10 $P_{N25/lac}$	0.903	10	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAAATTGTGAGCGGATAACA
11 $P_{N25/aUSR}$	1.301	11	GGCTGTGCGGCACGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT
12 $P_{N25/USR*}$	1.491	12	GGCTAAAAAACACGTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT
13 P_{con}	0.602	13	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCATAAGGAGGTGGATCCGGC
14 $P_{con/O3}$	1.072	14	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCAATTGTGAGCGGATAACAA
15 $P_{con/O5}$	1.173	15	GGATAACAATTTAGTTGACATTTTTAAGCTTGGCGGTTATAATGTTACCATAAGGAGGTGGGAATTCC
16 $P_{con/N25}$	1.398	16	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAATTTGAGAGAGGAGT
17 $P_{con/pex}$	1.204	17	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAAGGGTCGAGAGGAGT
18 $P_{con/anti}$	0.255	18	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATCCGGAATCCTCTTCCCG
19 P_L	1.724	19	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC
20 P_{L-8A}	1.672	20	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATAATGAGCACATCAGCAGGACGCACTGAC
21 P_{L-12T}	1.398	21	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATACTGAGCACATCAGCAGGACGCACTGAC
22 $P_{L/con}$	1.146	22	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATCAGCAGGACGCACTGAC
23 $P_{L/N25DSR}$	1.813	23	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATAAATTTGAGAGAGGAGT
24 $P_{L/con/N25DSR}$	1.778	24	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATAAATTTGAGAGAGGAGT
25 $P_{L/N25USR}$	1.763	25	CATAAAAAATTTATTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC
26 P_{LS1}	1.974[a]	26	TCCGTCTCGACGGGTTGACACAAAAGCCACAAGGGGTTATAATGAGCACATAAACTTGAGAGAGGAAT
27 P_{LS2}	1.968[a]	27	TGCGTATAGACAGTTTGACACAAAAGCCACAAGGTGTTATAATGAGCACATAAATTTGAGAGAGGAAT

[a] Predicted from the two dimensional PLS model.

P_{N25}, P_{con} and P_L were, furthermore, present also as 19 different site specifically altered or mixed variants. The sequences of these 25 promoters are presented together with the respective log *in vivo* promoter strength in Table 1. The logarithm of the promoter strength in P_{bla}-units is used in the modelling because of the large variation in strength.

DATA ANALYTICAL METHODS

The 25 promoters in Table 1 are parametrized in each position by the four descriptor variables earlier defined, giving a 25×272 matrix (**X**). The data matrix is not presented here, but can be regenerated from Table 1 using the descriptors from above. The parametrization enables the 25 sequences to be represented as a cluster of points in a 272 dimensional hyperspace. We are here aiming at identifying the structures in this space that are correlated with the promoter efficiency. Two multivariate data analytical methods that have been used successfully to solve analogous problems are principal components analysis (PCA) for graphical visualization of data structures, and partial least squares projections to latent structures (PLS) to establish quantitative relationships. Both of these methods are well explained in the literature, see e.g. (19–22). PCA decomposes a matrix **X** into means ($\bar{x}_k$), scores (t_{ia}), loadings (p_{ak}) and residuals (e_{ik}), according to;

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{A} t_{ia} \cdot p_{ak} + e_{ik}$$

Here the elements x_{ik} are the sequence descriptor variables with index i denoting promoters and k the sequence descriptors. The first principal component explains the largest part of the systematic variance of **X**, the next one the second largest part, and so on. One component (or dimension) of a PC-model thus consists of a score (column) vector t reflecting the systematics among the objects (here: sequences) times a row vector p characterizing the systematic behaviour (co-variance structures) of the sequence descriptor variables. Projection methods such as PCA and PLS summarizes most of the variation in a data matrix by a small number of derived variables, this regardless of the nature of the original variables i.e. whether they are continuous, discrete or a combination thereof (20). Moreover, as pointed out by Gower (23), the variables are combined in a way consistent with a similarity measure between objects (here: sequences) proportional to how many variables having the same values for any two objects. The number of statistically significant principal components (A) of a particular matrix is determined using cross-validation (24). This ensures that the model is not overfitted to the data, since this would severely hamper the predictive capability. The objective of cross-validation is thus to identify the number of components that provide an optimal balance between fit and predictive capabilities, in the present example A comes out to be two both for the PCA and the PLS models. Bivariate plots of the score values (t_{ia}) from different components provide projections of the data space, wherein systematic variance based patterns may be examined. Here, PC-A is used to obtain a graphical representation of the structural features of the synthetic sequences in relation to the original set of 25 promoters.

In the present application there are 272 variables (K) characterizing the structural features of 25 sequences (N). Methods like e.g. multiple and generalized regression can, therefore, not be applied since they require both that $N > K$ and that all the variables (K) are independent (i.e. orthogonal). PLS (21, 25) is used to correlate a single dependent variable y (or a matrix **Y**), to the variation in a predictor matrix **X**. PLS is a generalization of PCA where the components of **X** (t_a) are calculated so that they well approximate **X** and correlate well with y. Since PLS is a projection method, it can handle collinear data having many more variables (sequence descriptors, K) than objects (sequences, N), as long as the resulting components (A) are few compared to N. The result is a stable approximation of the correlation between **X** and y. The statistical significance of PLS models is also determined by cross-validation. In this paper PLS is used to relate the promoter efficiency variable (y) to the systematic variation in the promoter sequence matrix (**X**, the 25 parametrized 68 mers). The interpreted model is subsequently used to generate suggestions of sequences containing the essence of the structural features characteristic of strong promoters.

Choice of model

This QSAM was attempted using both PLS and a hetero-associative back-propagation NN, the results obtained were similar. However, we only present the results from the PLS QSAM since this method was found to be advantageous for a number of reasons, namely; **1**) PLS is more robust, since it does not require the proper initial setting of numerous variables (e.g. the number and size of hidden layer(s), choice of weight function(s), epoch length, etc.) which is required for the NN, **2**) PLS converges to the stable solution in a matter of seconds rather than hours, **3**) PLS proceeds in a fashion that allows the statistical significance of the model to be simultaneously evaluated. The risk of overfitting the model to the data is therefore minimized while the predictive capabilities are optimized, and **4**) the PLS weights are more readily interpreted both in terms of which is the optimal monomer in each position, but also if quantitative descriptors are used, the physico-chemical reason(s) **why** a certain monomer is preferred. The interpretation of the NN QSAM was not equally straightforward irrespective of the sequence descriptors used.

Results of the PLS QSAM

Two PLS dimensions, significant according to cross-validation, accounted for 27.5% of the systematic variance of **X** and explained 85% of the variance in the promoter efficiency variable (y). The first PLS dimension alone used 11.5% of **X** to explain 73% of y, the second added an additional 16% of **X** and 12% of y. These results refer to autoscaled data (i.e. each variable is scaled inversely proportional to its variance), the use of unscaled data, however, gives similar results. The PLS weights were subsequently transformed into PLS regression coefficients. The relative size of these coefficients over different positions indicate their relative importance to the promoter strength. To display the influence of each position we have in Fig. 1 summed the absolute values for each group of four coefficients and corrected this number for the degrees of freedom (DOF) for each position.

From figure 1 it can be seen that positions −135 to −33, −11 and +1 are constant for all 25 promoters in this example. These positions are probably important to the promoter strength, although the magnitude of their influence cannot be assessed from the present data. Among the varying positions we see that position −12 is the most important followed by positions 4 to 14 in the downstream region, positions around −8 to −10 and position

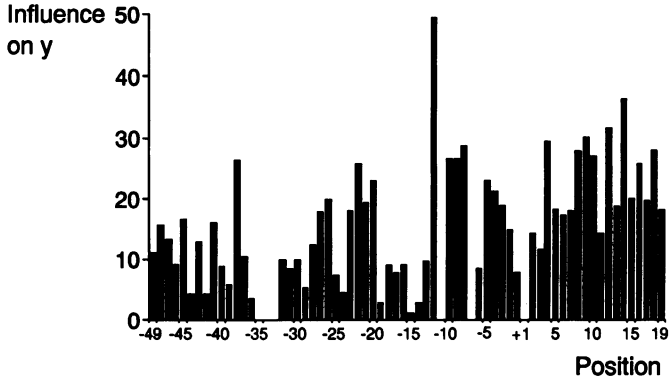

Figure 1. The influence on *in vivo* promoter strength of each position in the considered 68 bp fragment displayed as;

$$\text{Position influence} = \sum_{n=1}^{4} ((\,|c_n|\,)/\text{DOF}) \times 10^4$$

Where c are the PLS regression coefficients and DOF the degrees of freedom for each position, (the number of bases actually occurring in a particular position minus one). The correction for DOF results in that a relatively larger importance is given to the more conserved positions.

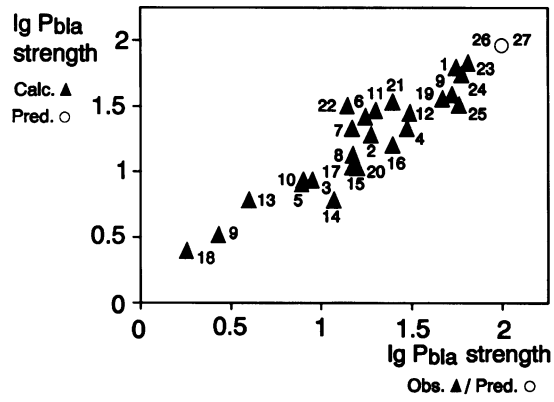

Figure 2. PLS correlation plot for the two dimensional model, showing the promoter strength calculated from the QSAM for the 25 training set objects plotted versus the corresponding literature data. The two structures suggested, P_{LS1} and P_{LS2} are included in the plot, using the predicted values on both axis. These are predicted to be the most potent promoters. Promoter strengths are given in logarithmic *bla*-units, numbers correspond to those in Table 1.

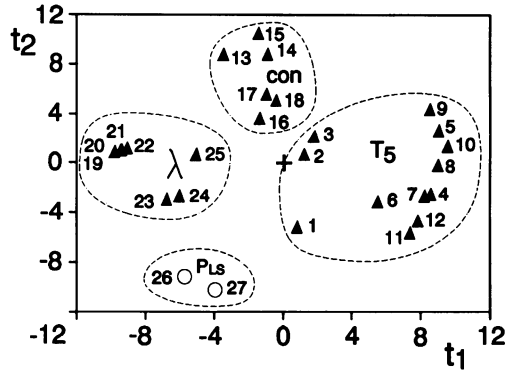

Figure 3. Score plot from the PCA on all 27 structures. The first component (t_1) describes 19% and the second (t_2) 11% of the systematics in the composition of the sequences. Numbers refers to those in Table 1.

−38. The relatively large influence of the downstream region seems to corroborate the observation that *in vivo* promoter strength is dependent on more than one functional parameter (15). The importance of this region may reflect the contributions to promoter strength from ease of initiation and/or the speed of RNAP promoter clearance.

The 272 PLS regression coefficients were then examined in detail, in groups of four corresponding to the descriptors for each of the 68 positions. For each of these groups the largest positive value indicates the 'preferred' base of that particular position. An entirely synthetic sequence was thus generated by selecting the most favourable base, with respect to the model. This was made both for the one and the two dimensional PLS models. The result was two strength-optimized 68 bp sequences denoted P_{LS1} and P_{LS2}. For the homologous and close to homologous positions in the −35, −10 and +1 regions the PLS sequences are determined by the bases having descriptors matching the corresponding column averages ($\bar{x}$). The P_{LS1} and P_{LS2} from the one and two dimensional QSAM were subsequently parametrized and reinserted into the model and their *in vivo* promoter strength was predicted, see Table 1 and Fig. 2. Promoter strength predictions from the NN QSAM were similar, (data not shown).

To visualise the sequence characteristics of these theoretical promoters in relation to those of the training set, all sequences were analyzed by PCA. The result was a significant two component model describing a total of 30% (19 and 11% respectively) of the systematic variance in the composition of the sequences. The scores from the two components are plotted in Fig. 3. The promoters of the training set are seen to be clustered according to their origin, (phage T5, phage lambda and constructs). The strength optimized constructs from the QSAM are positioned separately at the lower part of the projection. The four most potent training set sequences $P_{L/N25DSR}$, $P_{L/con/N25DSR}$ $P_{L/N25USR}$ and $P_{D/E20}$ are all seen to be situated in this region. The QSAM has thus pointed us further into the direction of strong promoters, outside the region defined by the original 25 sequences. Also in this analysis, the PC loadings show that the region downstream +1 is the main determinant of the patterns observed, (data not shown).

To validate the predictive capabilities of this QSAM it was subsequently decided to synthesize the promoters suggested by the model and determine their relative *in vivo* promoter strengths. An external reference set of six promoters ranging from weak to strong was kindly provided to us by H.Bujard and co-workers. However, since the reference promoters were found to be of different lengths, it was decided to include three versions of the reference promoters (P_{A1}, $P_{D/E20}$ and P_{Lcon}) having the same length and restriction sites at the positions corresponding to those of the test promoters, in order to obtain a better comparability.

EXPERIMENTAL

Vector construction

The P_{Lcon} insert was removed from pDS2 using *Xho*I/*Bam*HI and replaced by the *Xho*I/*Bam*HI cloning cassette fragment (containing an *Eco*RI site) from pGEM 7Zf+ (Promega), this

vector was designated pJJK1. A more thorough description of the test vector system pDS2 is given in (5).

Oligomer synthesis

The reference promoter $P_{D/E20}$ and the QSAM constructs P_{LS1} and P_{LS2} were synthesised as complementary 93 mer single stranded oligo-nucleotides, containing a *Bam*HI and a *Eco*RI site at the 5′ and 3′ ends respectively using a Beckman SM DNA synthesizer. The oligomers were collected and deprotected using the procedure recommended by the manufacturer, and thereafter purified according to (26). The 93 mers were purified on a non-denaturing 7% poly-acrylamide gel, annealed and trimmed to 75 mers by *Bam*HI/*Eco*RI treatment. The double stranded promoter fragments were further gel-purified as described above. Fragments were electro-eluted from the gel and cloned into the test vector pJJK1. Oligonucleotides complementary (40 bases into the transcript) to *bla* (27 mer) and *dhfr* (25 mer) were synthesised and radioactively labelled using terminal transferase (TdT) and α-^{32}P dATP (Amersham). Labelled oligonucleotides were purified on NENSORB columns (Du Pont NEN products), and used as probes in the dot-hybridization.

PCR amplification of reference promoters

Two reference promoters P_{Lcon} and P_{A1}, were obtained through PCR amplification generating 93 base pair fragments containing *Bam*HI and *Eco*RI sites at positions corresponding to the synthetic promoters. The double stranded fragments were enzymatically digested, gel-purified and cloned into pJJK1 as described above.

In vivo promoter strength determination

All test and reference promoter constructs were transformed into *E.coli* C600. Reference promoter constructs were the P_{Lcon} P_{A1}, P_{bla}, P_{con} and P_{N25} in plasmid pDS2 and p_{LA1sp} in plasmid pDS3. The DNA sequences of the positive transformants were confirmed by dideoxy sequencing (Pharmacia T7 sequencing kit). Bacterial cultures (50 ml) were grown as described in (5), 35 ml of the cell cultures were quickly chilled to 0°C, and the cells collected by centrifugation (1.5 min, 10krpm.). The remaining 15 ml was used to confirm sequence integrity. Cell pellets were quickly resuspended in RNA preparation buffer containing 7M urea, 3M LiCl, 0.2% SDS and 0.1% Antifoam A (Sigma), using a polytron, (10 sec, 24krpm), and RNA precipitated for 15h at 0°C. Total RNA was recovered by centrifugation (20 min, 15krpm), rapidly dissolved in 1×TE buffer containing 0.5% SDS and phenol extracted ×3 followed by CHISAM (chloroform:iso-amylalcohol 24:1 v/v) extraction ×1. The RNA was precipitated using two volumes ethanol and 0.3M KAc (pH 4.5). The quality of the RNA preparations were verified by northern analysis, and were found to be free from contaminating plasmid DNA.

Dot blot hybridization

Each RNA preparation was divided into 10 fractions (5 aliquots for repeated determinations of test and internal standard respectively). A dot hybridization analysis using Hybond N membranes (Amersham), was performed according to principles outlined in (27). The membranes were prehybridized for 15h at 65°C in 20 ml 6×SSC, 5× Denhardt solution, 1mM EDTA, 10mM Na_2HPO_4 (pH 7.0) and 250×10^{-6} g/ml of calf thymus DNA. The filters were hybridized (15h) in prehybridization solution containing 5×10^6 cpm/ml of radioactively labelled oligomer complementary to the *bla* and *dhfr* transcripts respectively. Dot blots were washed in 2×SSC, 0.1% SDS (15 min, 65°C). The amounts of mRNA expressed by test and reference promoters relative to the internal standard were subsequently determined by liquid scintillation of a defined minimum area of the dots.

Table 2.

Promoter	Measured strength (log *bla*-units)	Literature / Predicted[a] strength (log *bla*-units)
P_{LS1} (68 bp)	2.143	1.974[a]
P_{LS2} (68 bp)	2.127	1.968[a]
$P_{D/E20}$ (68 bp)	1.939	1.748
P_{A1} (68 bp)	1.556	1.643
P_{Lcon} (68 bp)	1.544	1.146
P_{A1}	1.707	1.643
P_{Lcon}	1.740	1.143
P_{bla}	1.041	-0.523
P_{con}	1.897	0.602
P_{N25}	1.799	1.477
P_{LA1sp}	1.107	1.929

Strength values for each promoter are averages based on five replicated measurements from a minimum of three different bacterial cultures. The experimental variability was generally between ±10−17% of the P_{bla}-strength (linear scale). The main source of this variation was, however, due to systematic differences in the dot-blot intensity of the internal standards.

RESULTS

When the relative promoter strengths of all 11 constructs were compared, the promoters generated from the QSAM were consistently found to be the strongest, see Table 2. On a linear scale the P_{LS1} was found to be 60% and the P_{LS2} 54% more efficient than the strongest reference. This is in good agreement with the predicted difference which is 68% and 66% respectively.

However, some discrepancies between the promoter efficiencies determined by us and those published earlier were found. The largest of these was displayed by the P_{LA1sp}, supposedly the strongest of the eight references, which in our system was found to be among the weaker. The reason for this is largely because the signal from the internal standard P_{bla} was substantially increased probably by the tendency of this promoter to initiate a backwards transcription into the *bla* region. This fact unfortunately, makes all comparisons with this reference promoter impossible. The second relevant difference is that the P_{con} was identified as being a relatively strong promoter in our system, the reason for this difference is not apparent. However, it may be noted that this promoter originally was tested in a different bacterial strain (*E.coli* M15). It was also observed that the strength of the longer versions of the P_{A1} and the P_{Lcon} are higher than those observed for the 68 bp versions, indicating either context effects from the sequence elements outside the 68 bp fragment, or an effect from the modification of the cloning

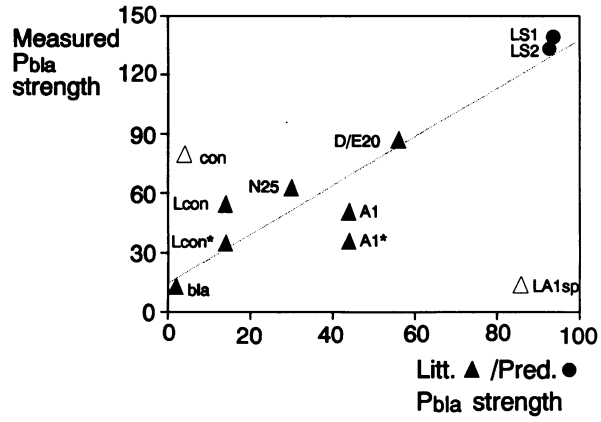

Figure 4. Plot of literature promoter strength (P_{bla}-units) for references and predicted values for P_{LS1} and P_{LS2} versus those experimentally determined. Promoters with a deviating behaviour are marked using open triangles. Promoter strength and names refer to Table 2. Asterisks refer to PCR versions of P_{A1} and P_{Lcon}.

cassette. The response in our system is also seen to be relatively higher than that earlier used. These trends were consistent although experiments were repeated. The sequence integrity of all constructs was also confirmed.

Some of the observed differences are probably due to the fact that for practical reasons, we used experimental conditions somewhat different from those earlier published. For example, we use labelled oligonucleotides (25 and 27 mers) as probes to detect the mRNA transcripts. The earlier determinations were made by *in vivo* incorporation of radioactive nucleotides into mRNA, detecting the *bla* and *dhfr* transcripts using larger M13 ssDNA probes (*bla* 752 and *dhfr* 672 bp respectively) (6). This difference in integrating the response from the P_{bla} internal standard (i.e. using a 25 mer or a 752 mer) may account for the fact that all promoter strength values for both reference and test promoters reported here are systematically higher compared to the earlier measurements. Apart from these inconsistencies it was noted that the P_{bla} also in our system is the weakest of the nine reference promoters and that $P_{D/E20}$ was judged to be the strongest. The strengths of the promoters P_{N25}, P_{Lcon} and P_{A1} are also found to fall between the strongest and the weakest of the reference set. The order for P_{Lcon} and P_{A1} is, however, reversed for the longer versions. The literature data for the reference promoters are plotted versus those experimentally determined in Fig. 4, together with the predicted activities for P_{LS1} and P_{LS2} and the corresponding experimentally determined values.

Considering that a relatively complicated experimental scheme has been transferred from one laboratory to another, the hierarchy of *in vivo* strength among the reference promoters was regarded to agree acceptably. It was therefore, concluded that the sequences generated from the present QSAM are strong promoters *in vivo*. They seem, in fact, to be stronger than any of the 25 promoters originally present in the training set, as predicted.

DISCUSSION

We demonstrate here the development of a nucleic acid QSAM for which the predictive capabilities have been experimentally verified. It is most encouraging that the principles of multivariate QSAM apply also here. The information that may be extracted from Table 1 is to some extent variance based; the fact that these 25 promoters differ considerably in strength cannot be explained in terms of homology. The differences between weak and potent promoters are, however, shown to be well modelled in terms of the systematic structural variation between the sequences. This variation is multi-positionally encoded and is hence analyzed accordingly. Three data analytical methods that have shown to be useful for this purpose are PCA, PLS and, to some extent, NN. The approach has general applicability for modelling the relationship between sequences (i.e. DNA, RNA and proteins) and their biological activity. The QSAMs can be interpreted in terms of important positions and regions. They can also be used to propose sequences for which the biological activity has been specifically altered.

The fact that promoter efficiency is affected by variable sequences, flanking more constant regions has earlier been demonstrated (5, 28). However, the identification and quantification of the systematic variations is usually not made. On the contrary, the homology approach aims to model variations in biological activity, in terms of a description of a local minimum of variance (the consensus regions). It has, however, been shown that consensus sequences mainly reflect the structures essential for efficient promoter recognition, but that the features determining the *in vivo* efficiency are less well described by homology (6). An additional drawback is that such models cannot be used to predict structures having a higher activity than those containing the 'optimal' consensus sequence. The term consensus thus is a statistical term that may not always be equated with strong or optimal (29). It is therefore suggested that homology based models are better suited for pattern recognition (i.e. classification) purposes. Knowledge about the homologous regions is, however, crucial in order to establish QSAMs.

Context (i.e. interaction) effects have also been shown to be important within consensus regions of bacterial promoters (30). The present models are linear. Interaction effects between bases are thus confounded (mixed) with the main effects from each position. The magnitudes of interaction effects may **only** be estimated if the alterations in a sequence are carried out according to an experimental design (i.e. D-optimal design) (31). The use of such designed sets of sequences in combination with multivariate data analysis will enable the magnitudes of both main and interaction effects for different positions to be estimated.

It was recently concluded that many subcellular processes will probably remain unintelligible unless properly quantified (32). The use of internal standards as well as external references in studies of functional parameters is therefore encouraged. This enables a better comparability of data from different sources and thereby greatly facilitate future QSAM development.

The present approach requires that the modelled sequences are of similar length. Moderate differences may be accommodated, but in the case of larger length deviations it is expected that QSAMs must be developed on a class by class basis or that sequence data is pre-treated by alignment independent sequence transformations. In the present example it was decided to include promoters of the same length since homogenous training sets are known to provide a more stable basis for predictions. We are currently developing quantitative monomer descriptors for bases occurring in DNA and mRNA, and also investigating the applicability of alignment independent transformations for sequence data. We hope to be able to report on these subjects in the near future.

ACKNOWLEDGEMENTS

This work was made possible through the Swedish Natural Science Council (NFR) grants; B-BU 4875-311, B-BU 2930 and K 3174-306. The generous support from Prof. Thomas Edlund and Prof. Glenn Björk, Department of Microbiology, University of Umeå is hereby gratefully acknowledged.

REFERENCES

1. Hawley, D.K. and McClure, R.W. (1983) *Nucleic Acids Res.*, **11**, 2237–2255.
2. Staden, R. (1984) *Nucleic Acids Res.*, **12**, 505–519.
3. Harley, C.B. and Reynolds, R.P. (1987) *Nucleic Acids Res.*, **15**, 2343–2361.
4. Galas, D.J., Eggert, M. and Waterman, M.S. (1985) *J. Mol. Biol.*, **186**, 117–128.
5. Cardon, L.R. and Stormo, G.D. (1992) *J. Mol. Biol.*, **223**, 159–170.
6. Mulligan, M.E., Hawley, D.K., Entriken, R. and McClure R.W. (1984) *Nucleic Acids Res.*, **12**, 789–800.
7. Kammerer, W., Deuschle, U., Gentz, R. and Bujard, H. (1986) *EMBO J.*, **5**, 2995–3000.
8. Knaus, R. and Bujard, H. (1990) Eckstein, F. and Lilley, D.M.J. (ed.), Nucleic Acids and Molecular Biology, Springer-Verlag, Berlin Heidelberg, Vol. 4, pp. 110–121.
9. van Heel, M. (1991) *J. Mol. Biol.*, **220**, 877–887.
10. Jonsson, J., Eriksson, L., Hellberg, S., Lindgren, F., Sjöström, M. and Wold, S. (1991) *Acta Chem. Scand.*, **45**, 186–192.
11. Brunak, S., Engelbrecht, J. and Knudsen, S. (1990) *Nucleic Acids Res.*, **18**, 4797–4801.
12. Demeler, B. and Zhou, G. (1991) *Nucleic Acids Res.*, **19**, 1593–1599.
13. O'Neill M.C. (1991) *Nucleic acids Res.*, **19**, 313–318.
14. Hellberg, S, Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S. and Andrews, P. (1991) *Int. J. Protein Res.*, **37**, 414–424.
15. Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström. M., Skagerberg, B. and Wikström, C. (1987) *Can. J. Chem.*, **65**, 1814–1820.
16. Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M. and Wold, S. (1989) *Quant. Struct.-Act. Relat.*, **8**, 204–209.
15. Brunner, M. and Bujard, H. (1987) *EMBO J.*, **6**, 3139–3144.
16. Knaus, R. and Bujard, H. (1988) *EMBO J.*, **7**, 2919–2923.
17. Lanzer, M. and Bujard, H. (1988) *Proc. Natl. Acad. Sci.*, **85**, 8973–8977.
18. Dobrynin, V.N., Kobrobko, V.G., Severtsova, I.V., Bystrov, N.S., Chuvpilo, S.A., Kolosov, M.N. and Shemyakin M.M. (1980) *Nucleic Acids Res.*, Symp. Ser. **7**, 365–376.
19. J.E. Jackson, (1991) *A Users Guide to Principal Components*, John Wiley & Sons, New York, USA.
20. Jolliffe, I.T. (1986) Principal Component Analysis, Springer Verlag, New York, pp. 200–204.
21. Martens H. and Naes T. (1991) *Multivariate Calibration*, John Wiley & Sons, Chichester, USA.
22. E.R. Malinowski (1991) *Factor Analysis in Chemistry*, 2nd ed., John Wiley & Sons, New York, USA, pp. 49–58 and 169–172.
23. Gower, J.C. (1966) Biometrica, 53, 325–338.
24. Wold, S. (1978) *Technometrics*, **20**, 397–405.
25. Dunn, III, W.J., Wold, S., Edlund, U., Hellberg, S. and Gasteiger, J., (1984) *Quant. Struct.-Act. Relat.* 131–137.
26. Sawadogo, M. and Van Dyke, M.W., (1991) *Nucleic Acids Res.*, **19**, 674.
27. Sambrook J, Fritsch, E.F. & Maniatis, T. (1989) Molecular Cloning A Laboratory Manual, Nolan, C. (ed.) 2nd ed., Cold Spring Harbor Laboratory Press, New York, pp. 7.54–55.
28. Jacquet, M.A. and Reiss, C. (1990) *Nucleic Acids Res.*, **18**, 1173–1143.
29. Cavener, D.R. and Ray, S.C. (1991) *Nucleic Acids Res.*, **19**, 3185–3192.
30. Graña, D., Gardella, T. and Susskind, M.M. (1988) *Genetics*, **120**, 319–327.
31. Box, G.E.P. and Draper, N.R. (1987) Empirical Model Building and Response Surfaces, John Wiley & Sons, New York, pp. 1–17.
32. Maddox, J. (1992) *Nature*, **355**, 201.