A Multivariate Representation and Analysis of DNA Sequence Data

Jörgen Jonsson, Lennart Eriksson, Sven Hellberg, Fredrik Lindgren, Michael Sjöström and Svante Wold

Research Group for Chemometrics, Department of Organic Chemistry, University of Umeå, S-901 87 Umeå, Sweden

Jonsson, J., Eriksson, L., Hellberg, S., Lindgren, F., Sjöström, M. and Wold, S., 1991. A Multivariate Representation and Analysis of DNA Sequence Data. – Acta Chem. Scand. 45: 186–192.

A new way to represent and analyze DNA sequence data is described. This approach complements methods currently used, in that it allows the systematic part of the variation between different sequences to be modeled. This can prove as informative as absence of variation (homology), which is the most widely used criterion for comparing sequence data. A multivariate sequence–activity model (SAM), for DNA-promoter sequences is presented, by which the relative promoter strength is modeled in terms of the primary DNA-sequence. The model is shown to have a good predictive capability. The coefficients from the model are interpreted, and used to design new structures predicted to be strong promoters in the system investigated. The approach described is also applicable to other kinds of sequence data, e.g. RNAs, proteins or peptides.

Recent advances in the field of genetic engineering have made it possible to clone and sequence virtually any piece of DNA of interest. This capability has, in turn, led to an almost explosive accumulation of DNA sequence data. This large source of information is often used as reference data for studies of new sequences for which the function is unknown. The goal of such studies is primarily to find sequences similar to that of interest, indicating the function of the protein encoded by the new sequence. Another use of this information is to compile large numbers of sequences of known function, in order to find regions of sequence homology i.e. 'consensus sequences'.[1,2] Such consensus sequences have, in turn, been used in attempts to forecast functional parameters (e.g. promoter strength) from structural parameters (i.e. DNA primary sequence) alone.[3] So far, however, such model sequences have been shown to be of limited predictive value.[4,5]

DNA sequences are usually represented using the four-letter code form (A C G T). However, when the biological activity of a set of related sequences is modeled, alternative ways of representing sequence data may be more practical. In addition, methods are needed that enable the major features of a set of sequences to be graphically displayed.

The first objective of this paper is to propose a parametrization of the four bases in the genetic code, which allows the systematic variation between different DNA, RNA or amino acid sequences to be visualized and analyzed in a new way. The second objective is to demonstrate that sequence–activity relationships, with predictive capabilities, can indeed be developed, provided that knowledge complementary to sequence homology is utilized. We emphasize that this information is inherent in sequence data. The problem is rather one of how to represent data and how to make this information numerically useful. The third objective is to show how the derived mathematical relationship between sequence variation and biological activity can be used to design new biomolecules where the biological activity is changed in a desired direction.

To exemplify how these objectives can be accomplished we have compiled from the literature DNA promoter sequences originating from *Escherichia coli* and coliphages, plus a number of artifical constructs. All these promoters have been consistently characterized with regard to their unregulated *in vivo* promoter strength by a standardized experimental system.[6]

Parametrization of DNA sequence data. In order to analyze sequence data, we need a numerical representation of the DNA (or RNA) monomer units, (the nucleosides A, C, G and T or U). Two different types of metric may be used for this representation; qualitative or quantitative. Qualitative parametrization corresponds to the use of descriptor variables of the indicator-type that, in a suitable way, describes the differences between the four nucleosides in DNA. Quantitative parametrization, on the other hand, uses continuous variables derived from measured physicochemical data of compounds of interest. We have derived both kinds of descriptor variables for the nucleosides, but since the resulting sequence–activity models (SAMs) for these two different ways of parametrization are very similar, we here present only the qualitative descriptors since they are easy to derive, few and conceptually simple.

To quantify four different objects in an unbiased way, it is sufficient to use three qualitative descriptor variables. By the term unbiased, we mean that no particular object(s) should be represented as being more similar or dissimilar with respect to any of the others. A geometric structure that meets this requirement is the tetrahedron. The corner

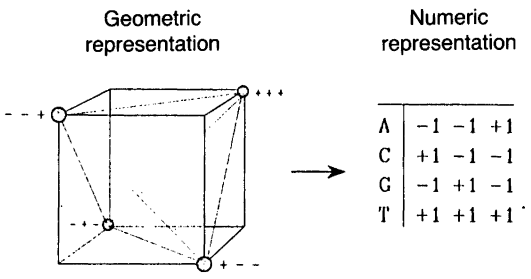

Fig. 1. Four diametrically opposed corners of a cube (circles) form a perfect tetrahedron. When placed in a three-dimensional coordinate system (origin in the center) each of the corners can be represented by a set of numerical coordinates. Since these descriptors are pure indicator variables it is of no importance which nucleoside is assigned to which corner of the tetrahedron.

coordinates for such a structure can thus be utilized as qualitative sequence descriptors. These coordinates can be conveniently generated by the use of a fractional factorial experimental design (FFD),[7] see Fig. 1. The objects (the nucleosides) are placed in four selected, diametrically opposed corners of a cube and all inter-object distances are identical. The structural variation in sequence data is thus described, simply by arranging the descriptor variables in accordance with the DNA sequences of interest. Each sequence will be transformed into a row-vector, and a number of sequences will form a sequence descriptor matrix.

Promoter sequence data. In prokaryotes, promoters consist of specific DNA sequences that govern the binding of the σ-unit of the RNA polymerase holoenzyme (RNAP), thereby punctuating the onset of transcription. There are large numbers of sequences known to act as promoters in *E. coli*; some of these originate from *E. coli* itself and a number of others from phages that act on this bacterium. Only a fraction of these sequences have, however, been characterized regarding *in vivo* promoter strength in a consistent manner. Efforts in this direction are presented in a series of articles by Bujard and co-workers.[4,6,8–10] A system that allows the *in vivo* efficiency of promoters to be determined, has been developed and subsequently used to determine functional parameters for some 28 promoters.

The 68 unit DNA sequences of these 28 promoters are presented together with their respective *in vivo* promoter strengths in Table 1. The promoter strength is given relative to the promoter for β-lactamase (P_{bla}), which is used as an internal standard. Monitoring of radioactively labeled mRNA expressed from the promoter under study, in relation to the standard, permits a relatively accurate determination of the promoter efficiency, unbiased by translational effects or gene dosage. This is just a brief presentation of a complex experimental system; a more thorough presentation can be found in Refs. 4 and 6.

The promoters in Table 1 are centered around the +1 base (start of transcription). Alternatively, it would have been possible to center the data around other positions known to be of relevance (e.g. the −35 or −10 region). The length of the spacer region, between the −10 and −35 region, does not differ with more than one base for this set of promoters. With greater differences we would expect that a number of sub-classes would have to be formed, and the data-analysis performed separately, on a class-by-class basis. The logarithm of the promoter strength in P_{bla} units is used in the SAMs; this transformation makes the data more normally distributed and thereby better suited to this kind of modeling.

Multivariate data and analysis. When the 28 promoters in Table 1 – each with 68 bases – are parametrized with the three descriptor variables defined in Fig. 1, the result is a 28×204 matrix. The data matrix is not presented here, but can be regenerated from Table 1 and Fig. 1. The information that can be retrieved from these data is, to some extent, multivariate; the fact that the promoters differ in strength cannot reasonably be explained by the features they have in common.

As a consequence of this parametrization, the unique sequence properties of each promoter can be represented by a single point in a 204-dimensional hyperspace. The promoters will thus form a cluster of 28 points in this space. Such a hyperspace has many properties in common with ordinary two- or three-dimensional spaces with which we are more familiar. There are angles, distances and planes but there is one crucial property that hyperspaces lack; they cannot easily be perceived by the human mind. There are, however, methods that allow hyperspaces to be studied in a rational way, thereby allowing us to identify systematic structures in such spaces, just as we do in ordinary two- and three-dimensional spaces. In this case it would be interesting to find sub-groups in the hyperspace and, perhaps even more interesting, to identify structures or trends in this space, correlated with the promoter efficiency.

Two multivariate data analytical methods that have successfully been used earlier to deal with such problems are principal components analysis (PCA)[11] for overview, and partial least-squares projections to latent structures (PLS)[12,13] to establish relationships. Both of these methods are extensively discussed in the literature and will therefore be only briefly described here. We here emphasize only that these methods can be used to analyze data matrices, such as the present one, that have many more columns (variables) than rows (objects).

The data matrix $\boldsymbol{X}$, is decomposed by PCA into means ($\bar{x}_k$), scores (t_{ia}), loadings (p_{ak}) and residuals (e_{ik}). In equation form this can be represented as eqn. (1). Here the elements x_{ik} are the sequence descriptor variables with index i denoting promoters and k their sequence descrip-

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

Table 1.

Promoter	Origin[a]	Sequence	Strength (log P_{bla} units)
		−49 −40 −30 −20 −10 +1 +10 +19	
1 D/E20	T5	ACTGCAAAAATAGTTTGACACCCTAGCCGATAGGCTTTAAGATGTACCCAGTTCGATGAGAGCGATAA	1.748
2 H207	T5	TTAAAAAATTCATTTGCTAAACGCTTCAAATTCTCGTATAATATACTTCATAAATTGATAAACAAAAA	1.740
3 N25	T5	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAATTTGAGAGAGGAGT	1.477
4 G25	T5	GAAAAATAAAATTCTTGATAAAATTTTCCAATACTATTATAATATTGTTATTAAAGAGGAGAAATTAA	1.278
5 J5	T5	TATAAAAACCGTTATTGACACAGGTGGAAATTTAGAATATACTGTTAGTAAACCTAATGGATCGACCT	0.954
6 A1	T7	ATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCATCGAGAGGGACACGGCGA	1.881
7 A2	T7	GAAAAACAGGTATTGACAACATGAAGTAACATGCAGTAAGATACAAATCGCTAGGTAACACTAGCAGC	1.301
8 A3	T7	TGAAACAAAACGGTTGACAACATGAAGTAAACACGGTACGATGTACCACATGAAACGACAGTGAGTCA	1.342
9 L	lam	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATCAGCAGGACGCACTGAC	1.568
10 con	ac	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCATAAGGAGGTGGATCCGGC	0.602
11 lac	coli	AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTGGTATGTTGTGTGGAATTGTGAGCGGATAACAA	0.756
12 lac/UV5	coli	AGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTGGTATAATGTGTGGAATTGTGAGCGGATAACAA	0.518
13 tacI	ac	TTCTGAAATGAGCTGTTGACAATTAATCATCGGCTCGTATAATGTGTGGAATTGTGAGCGGATAACAA	1.230
14 N25/03	ac	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAATTGTGAGCGGATAACAA	0.903
15 N25/pex	ac	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATAAAGGGTCGAGAAGAGT	1.176
16 N25/anti	ac	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCATCCGGAATCCTCTTCCCG	0.432
17 N25/lac	ac	CATAAAAAATTTATTTGCTTTCAGGAAAATTTTTCTGTATAATAGATTCAAATTGTGAGCGGATAACA	0.903
18 con/03	ac	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCAATTGTGAGCGGATAACAA	0.903
19 con/N25	ac	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAATTTGAGAGAGGAGT	1.398
20 con/pex	ac	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATAAAGGGTCGAGAGGAGT	1.204
21 con/anti	ac	ATTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGATTCATCCGGAATCCTCTTCCCG	0.255
22 con/D/E20	ac	TTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACCCAGTTCGATGAGAGCGATAA	1.114
23 con/trp	ac	TTCACCGTCGTTGTTGACATTTTTAAGCTTGGCGGTTATAATGGTACGCAAGTTCACGTAAAAAGGGT	0.903
24 L-8A	ac	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATAATGAGCACATCAGCAGGACGCACTGAC	1.672
25 L-12T	ac	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATACTGAGCACATCAGCAGGACGCACTGAC	1.398
26 L/con	ac	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATCAGCAGGACGCACTGAC	1.146
27 L/N25	ac	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTGATACTGAGCACATAAATTTGAGAGAGGAGT	1.813
28 L/con/N25	ac	TATCTCTGGCGGTGTTGACATAAATACCACTGGCGGTTATAATGAGCACATAAATTTGAGAGAGGAGT	1.813
		USR CORE DSR	
(a)		TCCGTAAAGAGAAG T CAAAATTCTCAACAGTCGT ATGCAGCCATAAATTTGAGAGAGGAGA	
(b)		tcTTGACA t t tg TATAAT CAT	
(c)		aAAAAa TTGCTa TATAAT TCAT TTGA	
(I)		TCCGTAAAGAGAAGttgaCAAAATTCTCAACAGTCGTtataATGCAGCCATAAATTTGAGAGAGGAGA	
(II)		TCCGTAAAGAGAAGttgctAAAATTCTCAACAGTCGTtataATGCAGCCATAAATTTGAGAGAGGAGA	
(III)		TCCGTAAAGAGAtcttgaCAAAATTCTCAACAtTtGTtataATGCAGCCATAAATTTGAGAGAGGAGA	

[a]ac = artificial constructs, lam = phage lambda. (a) Promoter structure deduced from SAM coefficients. (b) Consensus sequence proposed in Ref. 1. Bases that occur in at least 39 % in lower case, bases that are greater than 54 % conserved are in capitals. (c) Early T5 sequence elements from Ref. 6. (I)–(III) Promoter structures obtained by combination of fragments a–c, consensus 'parts' in lower-case characters.

tors. The principal components are calculated in the order in which the first component explains most of the variance of X, the next explains the second largest variance, and so on. The value of A, i.e. the number of statistically significant principal components (PCs) of a particular data matrix, is determined using cross-validation.[14] This significance test is applied in order to avoid overfitting i.e. an apparently good fit of the model to the data, but little predictive capability. In the present study PCA is used to make a graphical representation of sequence data. A plot of the values of t_{ia} for different A against each other provides a projection (or 'window') into the data space which displays the systematic patterns.

The PLS method is used to correlate the information in a matrix Y, (or a single variable y), with the variation in another matrix X. PLS is a generalization of PCA where separate components are calculated for each matrix, together with an inner relationship between the components (scores) of the two matrices. In this way a good approximation of the matrices X and Y is obtained, and at the same time a model with the maximum correlation between X and Y. The statistical significance of such PLS correlations is also tested by cross-validation. In this paper PLS is used to relate a promoter efficiency variable (y) to the variation in promoter sequence matrix (X, the parametrized promoter sequences).

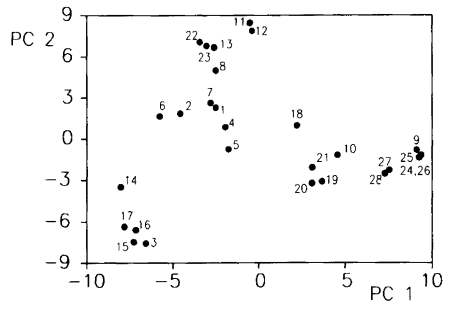

Fig. 2. A principal-component score plot, in which the PC scores (t_{ia}) of the two first components are plotted versus each other. Each sequence is represented by a filled circle. Similar sequences are found close to each other, dissimilar sequences are a long way apart. Numbering as in Table 1.

Graphical representation of sequence data. The PC analysis of the promoter descriptor matrix ($\boldsymbol{X}$), resulted in a model with four significant components (according to cross-validation). These four PCs describe 45 % of the variance in $\boldsymbol{X}$ (17, 11, 9.5 and 7.5 %, respectively). The PC scores (t_1–t_4) provide six different two-dimensional plots in which each promoter is represented by a point. In this paper we present and discuss only the first of these plots (Fig. 2) which also represents most of the information. This projection can be said to depict nearly a third (28 %) of the sequence variance among the compiled promoters.

Three clusters of promoters can be discerned in Fig. 2, the most prominent being that in the lower lefthand corner. The five objects in this cluster (Nos. 3 and 14–17), originate from the phage T7 promoter N_{25}, (No. 3). From the corresponding loadings (not shown here for reasons of space), it can be concluded that these objects are clustered by their relatively high AT content in the first two-thirds of the sequence (sometimes referred to as USR and core). Above this cluster a number of other phage promoters are seen, namely, E_{20}, H_{207}, G_{25} and J_5 (Nos. 1, 2, 4 and 5) from phage T5 and A_1–A_3 (Nos. 6–8) from phage T7.

In the central upper part of the figure are the two promoters lac and lacUV_5 (Nos. 11 and 12) originating from the *E. coli* lac operon. The tacI promoter (No. 13) which is a trp/lacUV_5 hybrid is also situated in this region. At the other end of PC_1, another phage promoter (No. 9) from phage lambda is found, together with some artificial constructs (Nos. 24–28 all originating from No.9). The phage lambda promoter is referred to in Ref. 10 as an 'alternative promoter structure.' This is corroborated by the PCA analysis, the main determinant being the relatively high GC content in the USR and core part of the sequence.

The result of this analysis on a set of numerically parametrized sequences is thus that similar (related) sequences form clusters of points situated closely together in the PC–score plots. Conversely, dissimilar sequences will be situated a long way apart. The corresponding PC loadings can then give information as to where in the sequences these similarities/dissimilarities are to be found.

Since the objective of this study is to show the utility of the approach, we refrain from making any further interpretations of this PCA model. A more thorough analysis of a larger set of promoter structures will be published elsewhere. The five remaining score plots and the corresponding loadings are available as supplementary material. We note that this kind of analysis can be performed on a ordinary microcomputer in a matter of minutes. The results are then displayed in a limited number of plots where similarities/dissimilarities in various regions are summarized. This alternative way of analyzing sequence data, may in certain cases, prove to be more informative than calculating a single 'homology score.' These two approaches should not, however, be considered to be mutually exclusive, but should rather be used simultaneously in order to gain more comprehensive knowledge concerning sequence data.

Development of a promoter SAM. The second analysis is aimed at deriving a sequence–activity model (SAM). The sequence data ($\boldsymbol{X}$) are related to the dependent promoter efficiency variable (y) by a PLS model, resulting in a statistically significant four-components model. The fourth component was of marginal significance according to the cross-validation criterion. Consequently, we henceforth discuss only the first three of these PLS components. With this model a total of 23 % of the variance in the sequence descriptor matrix ($\boldsymbol{X}$) accounts for 94 % (65 % cross-validated) of the variance in the *in vivo* promoter efficiency (y) among these 28 promoters. The observed versus the calculated promoter efficiency in log P_{bla} units is plotted in Fig. 3. The agreement is seen to be quite good.

To investigate further the predictive capability of the model we divided the data set arbitrarily into two subgroups, one with the sequences having odd numbers in Table 1 and the other containing the sequences with even numbers. Two different SAMs were subsequently derived, one in which the 'even' sequences served as a training set and the promoter strengths of the 'odd' sequences were predicted, and vice-versa.

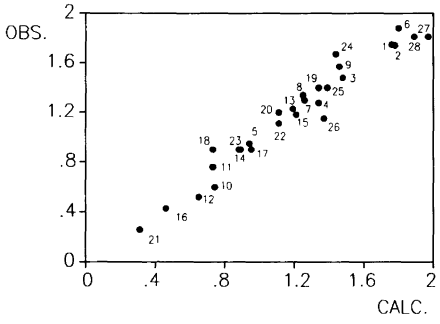

Fig. 3. Correlation plot in which the observed promoter strength (in log P_{bla} units) plotted versus the promoter strength calculated by means of the first SAM, based on all the 28 sequences. Notation as in Fig. 2.

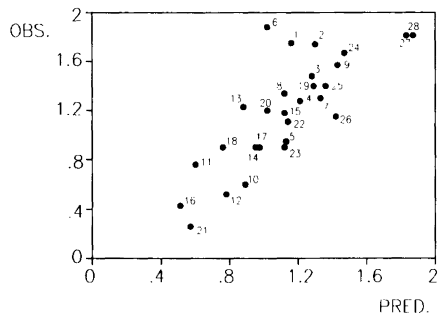

Fig. 4. Correlation plot in which *observed* promoter strength is plotted versus the promoter strength *predicted* by the two SAMs based on 'odd' and 'even' sequences from Table 1. Notation and scale as in Fig. 3.

This is a rather demanding test considering the limited size of the training sets; i.e. there are 4^{68} ($\approx 9^{40}$) different ways in which 4 nucleosides can vary in 68 positions. Even if only a very small fraction out of these 4^{68} possible structures can act as promoters *in vivo*, this 'limited' number can be anticipated to be rather impressive. Thus, when predicting the activity of 14 sequences from a training set of the same size, there is a substantial risk that structural features present in the test set sequences are not accounted for by the training set.

The result of this test is displayed in Fig. 4, where the predicted versus the observed promoter efficiency is plotted. We see that the majority of the predictions are acceptable, while they are not for some sequences (e.g. Nos. 1, 2 and 6). This may be explained by the limitations of the training sets, as discussed earlier. From the PLS loadings (not shown here for reasons of space) it can be concluded that it is the +1 to +20 region (denoted DSR) that has the largest influence on the relative *in vivo* promoter strength for this data set. This finding is in agreement with the results obtained by Kammerer *et al.*,[8] who have discussed the importance, with regard to promoter strength, of the sequences flanking the 'classical' −35 to +1 region. Also the early parts of the sequence (the USR) show significant PLS loadings. It may, therefore, prove useful to expand the definition of bacterial promoters also to encompass areas around the 50–70 base pairs commonly used, so that no information concerning the action of promoters is set aside.

From these results we conclude that it is indeed possible to establish multivariate SAMs that have predictive capabilities, provided that two criteria are met: (i) knowledge concerning systematic differences among 'similar' sequences must be utilized and (ii) it is crucial to have a balanced and well distributed training set, because of the large number of possible combinations. We have demonstrated that the first of these criteria may be fulfilled by a suitable data representation and the use of appropriate data-analytical methods. Unfortunately, it is not within our control to fulfill the second criterion. This does not imply that this is difficult, one way of accomplishing this will be discussed later. First, however, we wish to exemplify one possible use of the first model based on all 28 promoters.

Interpretation of the SAM. The result of the first SAM was that the dependent promoter efficiency variable (y) is expressed as a function of the independent sequence descriptors (X). In this process the relative importance and influence of each sequence descriptor on y, is recorded by the corresponding PLS loadings, one for each model dimension. The magnitude and sign of these coefficients can be used to determine which nucleoside is favorable in a certain position of a promoter sequence. These coefficients, in all 612 (3×204), are, as already mentioned, not shown here, but the principle of nucleoside selection based on the SAM coefficients is depicted in Fig. 5. Hence it is possible to generate (predict) a sequence that is 'optimal' in relation to the derived model parameters. Some of the positions with low variance (showing high homologies), i.e. the 'consensus sequences' in the −35 and −10 regions, will, as a consequence of the data-analytical method used, give small or ambiguous PLS loadings. This is not a problem, since the requirements for these regions have been extensively delineated by others.[1,15] A few other positions also showed ambiguous (but not always insignificant) loadings, indicating that the base selected for that particular position, in practice is less important for the promoter strength. In the case of such ambiguities, the base most closely corresponding to the signs of the loadings was chosen.

Thus, by combining the two complementary parts of knowledge (homology- and variation-based) it is possible to suggest promoter structures that are likely to have increased efficiency in initiating transcription in *E. coli*. We parametrized a number of these proposed structures and

	PLS-loadings		
Position 1	+65 +130 +45	→	T
Position 2	+8 −85 −21	→	C
Position 3	+11 −63 +39	→	A/C
Position 4	−22 +30 −4	→	G

Fig. 5. The principle of selecting an 'optimal' sequence from the loadings of the first PLS component for the positions −49 to −46 of the promoter sequence. The signs of the loadings are matched with the corresponding descriptors from Fig. 1. The choices for the first, second and fourth position are unequivocal. The sign combination for the third is not found in the descriptor table. Here two nucleosides A or C are equally favorable.

fitted them to the first SAM based on 28 sequences. These promoter structures were predicted to be some 30–50 P_{bla} units (28–47 %) more efficient than the strongest promoters in our training set. The sequence indicated by the SAM is listed at the end of Table 1 together with homology-based consensus sequences proposed in Refs. 1 and 5. We also present three suggested sequences likely to be strong promoters (I–III), but there are more that can be derived by the reader, by combining the fragments a–c.

It should be noted that the predictions of activities for these proposed sequences (I–III) are made by moderate extrapolations from the model and should as such be treated with some caution. Another important aspect is that the best sequences of the training set in fact promote transcription so efficiently that the forward rate constants for the complex formation between the promoter and the RNAP are thought to be at the limits that can be accounted for by ordinary diffusion encounters of the reactants.[6] In other words, these sequences have probably already been rather efficiently optimized by nature, and it is therefore anticipated that the sequences suggested by us may not be that much more efficient in practice, due to rate-limiting factors that cannot be mapped by the current data. However, we find it reasonable to believe that the structures proposed by us should act as strong functional promoters *in vivo* for *E. coli*.

Discussion

The objective of this last example is not to design a 'super promoter' but rather to demonstrate how this complementary piece of knowledge can make it possible to design rather complex biomolecules to have desired properties. We emphasize that this is a general approach applicable to most problems where the results are dependent on the joint influence of a large number of factors. We note that this is often the case in the areas of molecular and microbiology as well as in many other areas of research.

Many authors appreciate the multivariate nature of biological systems, but the approach to estimate these effects is often far from optimal. A common way of evaluating the importance of structural features is systematically to *c*hange *o*ne *s*equence element at a *t*ime (COST). This procedure, sometimes referred to as saturation mutagenesis,[16,17] tends to result in data where the information concerning interactions between different structural elements is inefficiently explored even though the number of experiments is large. An alternative way that has repeatedly been shown to be efficient for the evaluation of effects in complex systems is systematically to change many structural elements simultaneously according to an experimental design.[7,18,19] In this way both main effects and interactions are explored in a more efficient way by a limited number of experiments.

Some of the present promoter sequences have in fact been generated by Bujard and co-workers according to a plan that resembles a statistical design, in that relatively large fragments originating from both strong and weak promoters have been fused in order to generated hybrids having new combined features. This fact may contribute to the relatively good predictions for the two different test sets, discussed earlier.

From the results obtained here, it may be concluded that the only prerequisites for using multivariate methods to make graphical representations of sequence data are (1) a number of sequences (preferably > 10) that are structurally and/or functionally related and (2) a means of finding (or *a priori* knowledge of) an important position around which the sequences can be centered.

To develop multivariate SAMs we also need consistently measured biological responses. If a number of sequences are artificial constructs made to elucidate, alter or optimize some kind of mechanism, they should preferably be made according to an experimental design. By this procedure the 'structural space' will be spanned in a more efficient manner, thereby improving the predictive abilities of the models to be developed.

Concluding remarks. It is crucial further to develop methods that allow information to be extracted from sequence data, especially when considering the very intense research that is presently being planned and carried out in this area (e.g. the HUGO project). If too much emphasis is placed on homologies, large parts of vital information in sequence data may be neglected. It is probable that the research concerning DNA–protein, DNA–RNA and/or protein–protein interactions could benefit from the use of statistically designed sets of experiments combined with multivariate data analytical methods. In the view of the findings reported here, we are optimistic about the possibilities of establishing sound models that will facilitate the interpretation of data and also enable the design of biomolecules with desired properties. In our laboratory we are currently trying to refine the quantitative nucleoside descriptors while at the same time expanding them to encompass modified bases from DNA and RNA. We hope to be able to report on this in the near future.

Acknowledgements. Financial support from the Swedish Natural Science Research Council (NFR) and the *Kempe fund* at Umeå University is gratefully acknowledged.

References

1. Hawley, D. K. and McClure, W. R. *Nucl. Acid Res. 11* (1983) 2237.
2. Gren, E. J. *Biochemie 66* (1984) 1.
3. Mulligan, M. E., Hawley, D. K., Entriken, R. and McClure, W. R. *Nucl. Acid Res. 12* (1984) 789.
4. Geintz, R. and Bujard, H. *J. Bacteriol. 164* (1985) 70.
5. Galas, D. J., Eggert, M. and Waterman, M. S. *J. Mol. Biol. 186* (1985) 117.
6. Deuchle, U., Kammerer, W., Gentz, R. and Bujard, H. *EMBO J. 5* (1986) 2987.
7. Box, G. E. P., Hunter, W. G. and Hunter, J. S. *Statistics for Experimenters*, Wiley, New York 1978.

8. Kammerer, W., Deuchle, U., Gentz, R. and Bujard, H. *EMBO J. 5* (1986) 2995.
9. Brunner, M. and Bujard, H. *EMBO J. 6* (1987) 3139.
10. Knaus, R. and Bujard, H. *EMBO J. 7* (1988) 2919.
11. Wold, S., Esbensen, K. and Geladi, P. *Chemolab. 2* (1987) 37.
12. Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S. *J. Med. Chem. 30* (1987) 1126.
13. Eriksson, L., Jonsson, J., Sjöström, M. and Wold, S. *Chemolab. 7* (1989) 131.
14. Wold, S. *Technometrics 20* (1978) 379.
15. Dobrynin, V. N., Korobko, V. G., Severtsova, I. V., Bystrov, N. S., Chuvpilo, S. A. and Kolosov, M. N. *Nucl. Acid Res., Symp. Ser. 7* (1980) 365 and references therein.
16. Gaal, T., Barkei, J., Dickson, R. R., deBoer, H. A., deHaseth, P. L., Alavi, H. and Gourse, R. L. *J. Bacteriol. 171* (1989) 4852.
17. Dickson, R. R., Gaal, T., deBoer, H. A., deHaseth, P. L. and Gourse, R. L. *J. Bacteriol. 171* (1989) 4862.
18. Sjöström, M., Tosato, M. L., Eriksson, L., Lindgren, F., Hellberg, S., Jonsson, J., Marchini, S., Passerini, L., Pino, A., Skagerberg, B. and Wold, S. *Environ. Toxicol. Chem. 9* (1990) 265.
19. Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B. and Wold, S. *Int. J. Pept. Protein Res.* (1990). *In press.*

Received June 6, 1990.