Generating Programs for Predicting the Activity of Functional Sites

MICHAEL P. PONOMARENKO, ANNA N. KOLCHANOVA,
and NIKOLAY A. KOLCHANOV

ABSTRACT

The computer system ACTIVITY is intended for generating programs with which to predict the activity of functional sites by nucleotide sequences. ACTIVITY analyzes a basis set of nucleotide sequences with known activity. The novelty of this approach is that Zadeh’s fuzzy logic and decision-making theory have been employed for determining the best “sequence → activity” regression. The best one thus determined is then transformed into the text of a program with which the activity for any nucleotide sequence is to be predicted. Testing with independent data has proved this prediction reliable. We have compared our approach with the two commonly used on identical data sets to find the ACTIVITY-generated programs quite competitive.

Key words: nucleotide sequence; functional site; activity; prediction; program generation.

INTRODUCTION

There are more sites with known sequences than with known activity. Therefore, it is important that activity be predictable by sequence. Although a number of approaches exists, the problem as a whole is yet to be forestalled.

As is known, the activity of functional sites depends on nucleotide context. For example, the efficiency of su2-suppression of amber mutations in E. coli and S. typhimurium depends on which nucleotides are at positions −1 and +5 and which dinucleotide is at position +3 (Stormo et al., 1986). The frequency of mutations induced in the E. coli lacI gene by 2-aminopurine depends on which dinucleotide is at position −2 (Stormo et al., 1986). The constant, \( K_bK_2 \), of the rate of formation of “open-complexes” between RNA-polymerase and any E. coli promoter depends on the nucleotide context of the −35 and −10 boxes (Mulligan et al., 1984; Berg and von Hipple, 1987). The strength, \( P_{\text{bla}} \), of E. coli promoters depends on which nucleotides are at positions −38, −12, −10, −9, −8 and between +4 to +14 (Jonson et al., 1993).

The above examples demonstrate that different types of activity have different nucleotide contexts to depend on. That is why the basic problem is to answer the question as to which context feature is responsible for a particular type of activity. This can be done in a number of ways. For instance, the relationship between RNA-polymerase-binding activity, \( K_bK_2 \), of any E. coli promoter and the homology score of that promoter was uncovered by multiple alignment (Mulligan et al., 1984). Jonson et al. (1993) predicted the promoter strength, \( P_{\text{bla}} \), by mapping sequences of length \( L \) into a \( 4^L \)-bite space because neural networks are defined in a similar manner. By “generating and testing hypotheses”, Stormo et al.
(1986) found that some definite context features correlate with the efficiency of su2-suppression of amber mutations in E. coli, but his examples were few. For instance, a better prediction was achieved assuming the dinucleotide at position +3 than assuming the two nucleotides.

We propose an approach for determining the best "sequence → activity" regression on the basis of decision making theory (Fishburn, 1970) and Zadeh's (1965) fuzzy logic. In combination, these mathematical theories allow a much larger number of hypotheses found in Stormo et al. (1986) to be generated and tested, as automatic mode is now on.

The approach has been implemented as the computer system ACTIVITY. The system analyzes a basis set of sequences with known activity, determines the best "sequence → activity" regression, and saves it as the text of a computer program. The programs have proved competitive with neural networks (Jonson et al., 1993) and weight matrices (Stormo et al., 1986).

**SYSTEM**

ACTIVITY generates programs for predicting activity by nucleotide sequence.

ACTIVITY is to be fed with a basis set of sequences with known activity. An example of a basis set is presented in Table 1. The basis set AP20 consists of 15 sequences 11 bp in length centered at mutations induced by 2-aminopurine in the E. coli lacI gene (Miller et al., 1978). The number of such mutations on a logarithmic scale, ln[Nmut], is regarded as activity.

Fed with the basis set, ACTIVITY switches into automatic mode. Running ACTIVITY on a wide range of IBM/PC- and VAX-compatible computers indicates that processor time in the automatic mode only depends on computer capabilities and is about [24 hours] × r0/r (here: r and r0 are the respective efficiencies of the computer being run and a standard IBM PC/XT computer). The output is a computer program for the activity to be predicted by sequence.

The output for the basis set AP20 is shown in Figure 1. This is the text of a computer program in the "C" language. The program predicts the number of mutations that might be induced by 2-aminopurine in the center of a given sequence 11 bp in length under the specified experimental conditions (Coullondre et al., 1978). Once generated, the program does not require any ACTIVITY support.

Any such program predicts the activity, P(s), of a given nucleotide sequence, s, using single linear regression:

\[ P(s) = a_0 + a_1 \times X(s); \]

where: \(a_0\) and \(a_1\) are standard regression coefficients; \(X(s)\) is a context feature.

<table>
<thead>
<tr>
<th>Basis set, AP20</th>
<th>Control set, AP24</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Code</strong></td>
<td><strong>Sequence</strong></td>
</tr>
<tr>
<td>A6</td>
<td>TGAACACGGCC</td>
</tr>
<tr>
<td>O35</td>
<td>TGCTGCAACTC</td>
</tr>
<tr>
<td>A34</td>
<td>AGGCCACGGGC</td>
</tr>
<tr>
<td>O17</td>
<td>TCGCCGACAAG</td>
</tr>
<tr>
<td>A19</td>
<td>GTCACCAGCAGA</td>
</tr>
<tr>
<td>A21</td>
<td>TGTCGCAGCAG</td>
</tr>
<tr>
<td>O10</td>
<td>CACAAACACTG</td>
</tr>
<tr>
<td>O28</td>
<td>TCAACAAACC</td>
</tr>
<tr>
<td>A16</td>
<td>CGTACCAGCAG</td>
</tr>
<tr>
<td>A31</td>
<td>TCAACAGGGT</td>
</tr>
<tr>
<td>A33</td>
<td>TCTCTCAGGGG</td>
</tr>
<tr>
<td>O13</td>
<td>CGAATCAAATG</td>
</tr>
<tr>
<td>A5</td>
<td>CTATCAGACC</td>
</tr>
</tbody>
</table>

*Nucleotide sequences from E. coli lacI gene (Miller et al., 1978) and the number of mutations induced by 2-aminopurine on a logarithmic scale ln[Nmut] (Coullondre et al., 1978).*
Simple linear regression (1) was our preference by the following reasons:

1. This is the commonly accepted empirical estimation.
2. This is the simplest mathematical model for biological interpretation.
3. This is the simplest of all "sequence → activity" regressions possible.
4. This demands a minimum body of experimental data necessary for optimization.

Of course, this choice scarcely lets the investigator guess the useful context feature X. ACTIVITY removes the problem by "generating and testing" as many xs as the computer can afford. And this is novel.

In terms of the present version of ACTIVITY, the context feature X(s) is defined as the weighted number of k-mers, z_1z_2...z_k, contained in the sequence s = s_0s_1s_2...s_N:

\[ X_{Z_w}(s) = \sum_{i=a,b-k+1} \delta_z(s_i s_{i+1} ... s_{i+k-1}) \times w(i) \]

where: \( \delta_z(s_i s_{i+1} ... s_{i+k-1}) = 1 \) if \( z_j = s_{i+j-1} \), and \( \delta_z(s_i s_{i+1} ... s_{i+k-1}) = 0 \) if otherwise; \( w(i) \) is the weight of the \( i \)th position, and \( k \ll b - a + 1 \).

In Eq. (2), all possible k-mers, \( z = z_1z_2...z_k \), 1 to 4 bp in length encoded as \( \{ A, T, G, C, W = A/T, R = A/G, M = A/C, K = T/G, Y = T/C, S = G/C, B = T/G/C, V = A/G/C, H = A/T/C, D = A/T/G, N = A/T/G/C \} \) are used. Their total is \( 15 \times 15 \times 15 \times 14 = 47250 \).

Each function \( w(i) \) has a peak defined by indices \( \{ I_1, I_2, I_3 \} \). The peak may be either a maximum \( (I_1 = 0) \) or a minimum \( (I_1 = 1) \). The peak is located by \( I_2 \) that ranges between 1 and 15, peak width is governed by \( I_3 \) that ranges between 1 and 6. All possible combinations of the indices define \( 2 \times 15 \times 6 = 180 \) such functions. Thus the effects of localization of the k-mers \( z_1z_2...z_k \) on activity are simulated.

Each context feature \( X_{Z_w} \) calculated by Eq. (2) is defined by one of \( 47,250 \) k-mers, \( z \), and one of 180 functions, \( w(i) \). Their total number is \( 47,250 \times 180 = 8,505,000 \).

Eq. (2) is handled by the computer procedure WeightSum shown in Figure 1. The text of the procedure is written into the header file "activity.h" (Fig. 1). By the way, to use any ACTIVITY-generated program, it is only necessary to have this file attached.

How do Eqs. (1) and (2) handle an arbitrary sequence? For instance, the sequence is TGAACCAGGCC starting at position -5 (Table 1, sequence A6). For the program shown in Figure 1, the following equations

![Fig. 2](image-url)
hold true: $k = 3, z_1z_2z_3 = \text{SNR}$, and $w(i)$ is as in Figure 2. Trimmers SNR are at positions $-4, -1, 0$ of this sequence. Their $w(i)$ are 0.1, 0.5, and 0.7, respectively. Subject to Eq. (2), $X_{\text{SNR};w} = 0.1 + 0.5 + 0.7 = 1.3$, and subject to Eq. (1), $\ln[N_{\text{mut}}] = -2.156 + 4.299 \times 1.3 = 3.34$.

Obviously, the programs are manageable.

**METHOD**

It is postulated that context features are of importance unless otherwise proved. At any context feature $X_{z;w}$ calculated by Eq. (2), the $k$-mer $z = z_1z_2\ldots z_k$ simulates the optimum subsequence of a given functional site, and the function $w(i)$ simulates the optimum localization of the subsequence within the site.

In the present example (Fig. 1), $z = \text{SNR}$ and $w(i)$ reaches its maximum over position 0 (Fig. 2), and so was the best $X_{z;w}$ for prediction of the number of mutations, $\ln[N_{\text{mut}}]$, in this position. This implies that the more trimers SNR near the mutation point, the more mutations induced by 2-aminopurine. That was the ACTIVITY decision made on the basis set AP20 (Table 1).

Why did ACTIVITY decide on the $X_{\text{SNR};w}$ while there was a total of $8,505,000$ alike?

For the basis set of sequences with known activity $P$, to each context feature $X$ an estimate $u(X \rightarrow P)$ is assigned. In decision making theory (Fishburn, 1970), this estimate is termed “utility” because the higher $u(X \rightarrow P)$, the better $X$ suits $P$.

In the present example (Fig. 1), each of all $8,505,000$ possible $X_{z;w}$ was tested on the basis set AP20 for the significance of prediction “$X_{z;w} \rightarrow \ln[N_{\text{mut}}]$,” and to each the respective utility $u(X_{z;w} \rightarrow \ln[N_{\text{mut}}])$ was assigned. The value $u(X_{\text{SNR};w} \rightarrow \ln[N_{\text{mut}}]) = 0.266$ was the highest and so ACTIVITY made its choice. The algorithm of calculating utility is what makes ACTIVITY different from any other activity prediction device.

The algorithm is intended for optimizing simple linear regression (1), $X(s) \rightarrow P(s)$, on a given basis set of sequences, $s$, with known activity, $P$, and testing for the significance of the equation $P(s) = P$ on a larger number of different data sets. Seven such data sets are included in the algorithm:

1. the given basis set;
2. half the basis set with low $P$;
3. half the basis set in the vicinity of mean $P$;
4. half the basis set with high $P$;
5. half the basis set with low $X(s)$;
6. half the basis set in the vicinity of mean $X(s)$; and
7. half the basis set with high $X(s)$.

On each of these, the following eleven mathematical properties of the equation $P(s) = P$ are tested for significance using the relevant statistical criteria (Hajek and Sidak, 1967; Forster and Ronr, 1979; Lehman, 1959; Likes and Laga, 1978):

1. linear correlation between $P(s)$ and $P$;
2. Spearman’s rank correlation between $P(s)$ and $P$;
3. Kendall’s rank correlation between $P(s)$ and $P$;
4. the difference of $P(s)$ and mean $P(s)$ has the same sign as that of $P$ and mean $P$;
5. mean $P(s)$ is equal to mean $P$;
6. the variance of $P(s)$ is equal to the variance of $P$;
7. the difference of $P(s)$ and $P$ is a Gaussian distribution;
8. no difference of $P(s)$ and $P$ is dependent;
9. there are as many positive differences of $P(s)$ and $P$ as the negative;
10. $P$ is distributed uniformly; and
11. $P(s)$ is distributed uniformly.

These properties supplement one another, hence simple linear regression is relevant. For instance, one pair, $P$ and $P(s)$, may have departed considerably from the others, yet the coefficient of linear correlation can still strike a high value (property 1 above). In this “misleading” situation, the two rank correlations (properties 2 and 3 above) reject the usefulness of any regression.
ACTIVITY OF FUNCTIONAL SITES

The algorithm performs three regular program generation operations, namely rewriting, compiling, and testing, and additionally, decision making.

**Operation 1 (rewriting).** One of all 8,505,000 possible context features \( X_{z;w} \) is fixed by defining one of 47,250 possible \( k \)-mers, \( z = z_1 z_2 \ldots z_k \), and one of 180 possible functions \( w(i) \). This makes Eq. (2) set for application.

**Operation 2 (compiling).** Subject to Eq. (2), for every sequence \( s \) in the basis set, the value \( X_{z;w}(s) \) is calculated. Using these pairs \( \{X_{z;w}(s) \rightarrow P\} \) as commonly accepted (Forster and Ronr, 1979), the regression coefficients \( a_0 \) and \( a_1 \) are calculated. This makes Eq. (1) and the program ready for use.

**Operation 3 (testing).** The above seven data sets are compiled and, accordingly, seven data sets \( \{P_{z;w}(s) \rightarrow P\} \) are generated. On each of the latter, each of the above eleven mathematical properties are tested for significance. The total number of the significance levels is \( 11 \times 7 = 77 \).

**Operation 4 (decision making).** Each of the 77 significance levels are decoded by Zadeh's fuzzy logic (1965). A positive mark from 0 to 1 is assigned to each significant property, and a negative mark from \(-1\) to 0 to each insignificant. In terms of decision making theory (Fishburn, 1970), each of these 77 marks estimates on a common scale utility of the simple linear regression \( X_{z;w}(s) \rightarrow P(s) \). Decision on the regression is presented as the mean of the 77 marks, \( u(X_{z;w} \rightarrow P) \) so that:

1. \( u(X_{z;w} \rightarrow P) < 0 \) implies that the regression \( X_{z;w}(s) \rightarrow P(s) \) is groundless;
2. \( u(X_{z;w} \rightarrow P) > u(X_{q;v} \rightarrow P) \) implies that the regression \( X_{z;w}(s) \rightarrow P(s) \) is better motivated than \( X_{q;v}(s) \rightarrow P(s) \). Thus the highest \( u(X_{z;w} \rightarrow P) \) pinpoints the best \( X_{z;w} \).

**RESULTS AND DISCUSSION**

Since weight matrices (Stormo et al., 1986) has an impact on ACTIVITY, comparing the two approaches is a matter of course.

Sequences of the \( E. coli lacI \) gene with \( \ln[N_{mut}] \) mutations induced by 2-aminopurine (Miller et al., 1978; Coullondre et al., 1978) were examined as described at length in Sections SYSTEM and METHOD.

The program given in Figure 1 was tested on the control set AP2\(_a\) (Table 1). Prediction of the mutation number is presented in Figure 3. The coefficient of linear correlation between the predicted and known activities is \( r^2 = 0.740 \) (\( \alpha < 10^{-3} \)).

A very close estimate \( (r^2 = 0.749) \) had been obtained using weight matrices (Stormo et al., 1986). Thus, ACTIVITY is at least as reliable as weight matrices.

Neural networks are now widely used for activity prediction, and so we have compared ACTIVITY with one of them.

![FIG. 3. The results of the program given in Figure 1 on the control set AP2\(_a\) (Table 1).](image-url)
TABLE 2. STRENGTH OF E. COLI promoters

<table>
<thead>
<tr>
<th>Basis set, Ec₀</th>
<th>Control set, Ec₉</th>
</tr>
</thead>
<tbody>
<tr>
<td>L/N25DSR</td>
<td>1.183</td>
</tr>
<tr>
<td>D/E20</td>
<td>1.748</td>
</tr>
<tr>
<td>L</td>
<td>1.724</td>
</tr>
<tr>
<td>N25</td>
<td>1.477</td>
</tr>
<tr>
<td>G25</td>
<td>1.278</td>
</tr>
<tr>
<td>J5</td>
<td>0.954</td>
</tr>
<tr>
<td>N25/lac</td>
<td>0.903</td>
</tr>
<tr>
<td>con</td>
<td>0.602</td>
</tr>
<tr>
<td>con/anti</td>
<td>0.255</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The strength of E. coli promoters expressed in log[P₁₆] (Jonson et al., 1993). The nucleotide sequences are available in Table 1 by Jonson et al. (1993).

Sequences of 27 E. coli promoters (positions between −49 and +19) with known strength − log[P₁₆] (Jonson et al., 1993) were examined. These experimental data are presented in Table 2. The basis set Ec₀ of 9 promoters was entered into ACTIVITY, the other 18 promoters were the control set Ec₉.

On the basis set Ec₀, ACTIVITY generated a program which was defined by $k = 3$, $z_{1}z_{2}z_{3} = ASM$, and the function $w(i)$ as in Figure 4a, the coefficients $a_0 = 0.016$ and $a_1 = 0.740$, and utility $u(X_{ASM,w} \rightarrow \log[P₁₆]) = 0.589$. As can be seen from Figure 4, the trimers ASM between positions −12 and +14 were found to be responsible for strength − log[P₁₆]. This was tested on the control set Ec₉ (Table 2). The results obtained are presented in Figure 4b. The correlation between the predicted and known strength − log[P₁₆] was $r^2 = 0.741$.

These experimental data had been analyzed using neural networks (Jonson et al., 1993) and predictions were reliable. According to neural networks data, strength − log[P₁₆] depends on what sort of nucleotides are at positions −38, −12, −10, −9, −8 and between positions +4 and +14. By Fisher’s precise criterion (Lehman, 1959), ACTIVITY did as much ($\alpha < 10^{-3}$).

![FIG. 4. The “sequence → activity” regression for the E. coli promoter strength, − log[P₁₆], determined by ACTIVITY on the basis set Ec₀ (Table 2): (a) the weight function w(i) used in Eq. 2; (b) prediction on the control set Ec₉ (Table 2).]
ACTIVITY OF FUNCTIONAL SITES

Table 3. Nucleotide Sequences and Cleavage Frequencies

<table>
<thead>
<tr>
<th>Basis set, E2A₀</th>
<th>Control set, E2A₉</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence</strong></td>
<td><strong>ln[Pᵢ]</strong></td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>0.28</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTCGAGG</td>
<td>0.01</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>0.00</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>-0.01</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>-0.05</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>-0.15</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>-0.40</td>
</tr>
<tr>
<td>TTCTACCGGATCTTTTTTTTCGAGG</td>
<td>-0.42</td>
</tr>
<tr>
<td>TTCTACCGGATTTTTTTTTTCGAGG</td>
<td>-3.91</td>
</tr>
</tbody>
</table>

The nucleotide sequences and cleavage frequencies, ln[Pᵢ], for the downstream region of the pre-mRNA 3'-cleavage/poly(A) site E2A of the adenovirus (McDevitt et al., 1986).

ACTIVITY also was tried on experimental data which had not been analyzed before.

The data are presented in Table 3: 16 sequences 25 bp in length located between positions +1 and +25 relative to the cleavage point of the 3'-cleavage/poly(A) site E2A of the adenovirus pre-mRNA with known cleavage frequency ln[Pᵢ] (McDevitt et al., 1986).

These data were grouped into two sets, the basis set E2A₀ with 10 sequences and the control set E2A₉ with the other 6. On the basis set E2A₀, ACTIVITY generated a program defined by k = 4, z₁₂₃₄ = HNAWK, and the function w(i) as in Figure 5a, coefficients α₀ = −3.121 and α₁ = 1.370, and utility u(XASM,w) → ln[Pᵢ] = 0.356. This implies that the tetramers HNAWK between positions +3 and +15 were found responsible for the cleavage frequency (Fig. 5a).

Testing was performed on the control set E2A₉ (Table 3). The results obtained are in Figure 5b. The predictions fit in with the experimental data, r² = 0.831 (α < 0.025). Thus, the well-known T/G-rich downstream element located between positions +13 and +20 of this site was successfully detected by ACTIVITY (Fig. 5).

A program was generated for a nucleotide substitution experiment (Table 3). What would the prediction of cleavage frequency be like if the program were tried on insertions? With insertion, position numbers of the downstream element become higher. As can be seen from Figure 5a, the weight function w(i) has a maximum at position 10 after which it decreases smoothly. Therefore, the weights w(i) of the tetramers HNAWK of this element decrease and so does cleavage frequency. McDevitt et al. (1986) detected a gradual decrease in cleavage frequency when insertions took place between the downstream element and the cleavage point.

![Graph](a) ![Graph](b)

**FIG. 5.** The "sequence → activity" regression for the cleavage frequency, ln[Pᵢ], of the pre-mRNA 3'-cleavage/poly(A) site, E2A, of the adenovirus determined by ACTIVITY on the basis set E2A₀ (Table 3): (a) the weight function w(i) used in Eq. 2; (b) prediction on the control set E2A₉ (Table 3).
Taken together, the results obtained allow the conclusion that ACTIVITY provides novel research capabilities of automatic selection of context features suitable for predicting functional activity by nucleotide sequences.

ACKNOWLEDGMENTS

We are thankful to V. Filonenko for translating this manuscript into English. This research was supported by the Russian Ministry of Sciences, Education and Technique Policy, the Office of Health and Environmental Research of the Department of Energy, U.S.A. (grant #OR00033-93CIS002) and the Russian Fundamental Research Foundation.

REFERENCES


Address reprint requests to:
Michael P. Ponomarenko
Institute of Cytology and Genetics,
Siberian Branch
Russian Academy of Sciences
10 Lavrentyeva Ave.
Novosibirsk, 630090, Russian
pon@cgi.nsk.su

Received for publication August 3, 1995; accepted December 4, 1996.