GENOMIC SIGNAL PROCESSING

Lecture 3

Nonlinear Prediction of Gene Expressions from Microarray Measurements

1. Gene prediction problem
2. Comparison of prediction capabilities of several predictors
3. Search of all possible groupings of determination factors. Limiting the space of search.
4. Predicting a target gene using several hundred predictor genes

Gene prediction problem

- Postulate a parametric model for the dependency $g_\ell = f(g_1, \ldots, g_n)$.
- Fit the parameters to the available data.
- Find the codelength for sending the parameters and the residuals.

The MDL principle considers the following axiom as basis to the theory of modeling:

- Given the data and the model class, select the model in the model class which achieves the shortest codelength for the data and the model.
- The codelength necessary for encoding the data may be given by a two part code: first encode the model parameters, second encode the model residuals.

Data available

- The available data is organized in a matrix where row i is the set of measurements at instant i, the column j ($j = 1, \ldots, n$) represents gene g_j activity. The entries $M(i,j)$ take values in the set $\{0, 1, 2\}$.

We consider here a comparison of three predictors under MDL criterion, by using the experimental data from [Kim et al.2000], where M is a 30 × 14 matrix.

RCH1	BCL3	FRA1	REL-B	ATF3	IAP-1	PC-1	MBP-1	SSAT	MDM2	p21	p53	AHA	OHO
-1	1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	1	0	0	0	0	1	1	1	1	0
-1	0	0	1	1	0	1	0	0	1	1	1	1	1
0	0	1	0	1	0	0	0	0	0	1	1	1	0
-1	0	0	1	1	1	1	1	0	1	1	1	1	1
0	0	0	0	1	0	0	0	0	1	1	1	1	0
0	0	0	0	0	0	0	0	0	1	1	1	0	0
0	0	0	0	1	0	0	0	0	0	1	1	1	0
0	0	0	0	1	0	0	0	0	0	1	1	1	0
-1	0	1	1	0	0	0	0	0	1	1	1	0	1
0	0	1	0	1	0	0	0	0	1	1	1	1	0
0	0	1	1	1	0	0	0	0	1	1	1	1	0
0	1	0	1	1	1	1	0	0	1	1	1	1	0
0	0	0	0	1	0	0	0	0	0	1	1	1	0
0	0	0	0	1	0	0	0	0	0	1	1	1	0
-1	1	1	1	1	0	1	0	0	0	0	-1	-1	0
0	0	0	0	1	0	0	0	0	0	0	-1	0	0
-1	1	0	1	1	0	1	0	1	0	1	-1	-1	-1
0	0	1	0	1	0	0	0	0	1	1	-1	0	1
0	0	0	0	0	0	0	0	0	0	0	-1	0	0
0	0	0	0	1	0	0	0	0	0	0	-1	0	0
0	0	0	1	0	0	1	0	0	0	0	-1	0	0
0	0	0	0	1	0	0	0	0	0	1	-1	0	1
0	0	0	0	1	0	1	0	0	0	1	-1	0	1
-1	1	0	1	0	1	1	0	0	0	0	-1	-1	0
-1	0	0	0	1	0	0	0	0	0	1	-1	-1	1
-1	0	0	0	1	0	0	0	0	0	1	-1	-1	1
0	0	0	1	0	0	0	0	0	0	1	-1	0	-1
0	0	0	0	1	0	0	0	0	0	1	-1	0	1
0	0	0	0	1	0	0	0	0	0	1	-1	0	1

Comparison of prediction capabilities of several predictors

Hard Boolean predictor. Predictor specification and design

Denote $\underline{x} = [x_1, \ldots, x_n]$ the prediction window, having dimension n, and $x_i \in \{0, 1, 2\}$. Define the thresholded vectors $\underline{x}^{b_1} = [x_1^{b_1}, \ldots, x_n^{b_1}]$, and $\underline{x}^{b_2} = [x_1^{b_2}, \ldots, x_n^{b_2}]$, with

$$x_i^{b_1} = \begin{cases} 0 & if \quad x_i = 0 \\ 1 & if \quad x_i \geq 1 \end{cases} \tag{1}$$

$$x_i^{b_2} = \begin{cases} 0 & if \quad x_i \leq 1 \\ 1 & if \quad x_i = 2 \end{cases} \tag{2}$$

The prediction is defined as

$$\hat{y} = f(\underline{x}^{b_1}) + f(\underline{x}^{b_2}) \tag{3}$$

where $f(\cdot)$ is a Boolean function with n variables.

To design the Boolean predictor we quantize to two intervals the conditional expectation in the binary planes:

$$f^*(\underline{x}^b) = \begin{cases} 0 & if \quad \frac{1}{2}\left[E(y^{b_1}|\underline{x}^{b_1} = \underline{x}^b) + E(y^{b_2}|\underline{x}^{b_2} = \underline{x}^b)\right] \leq 0.5 \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

Hard Boolean predictor. Description length

- The model cost for encoding the function $f^*(\cdot)$ is

$$C_M(n) = 2^n \text{ bits} \tag{5}$$

  if we assume a uniform apriori distribution over all Boolean functions.

- The prediction errors can be encoded by favoring the string of all 0's (perfect prediction) as follows:

  - specify the location of error (a 5 bit pointer is needed for each incorrect predicted value in the string of 30 values)

  - specify the correct value of the data at that location (another 2 bits).

- Therefore if there are N_e nonzero prediction errors, the overall coding length will be the model length plus the prediction error length:

$$L(n) \;=\; C_M(n) + 7N_e = 2^n + 7N_e \text{ bits} \tag{6}$$

Ternary predictor

Predictor specification and design

- The optimal Ternary predictor is found by quantizing to three intervals the conditional expectation:

$$\hat{y} = h^*(\underline{x}) = \begin{cases} 0 & if \quad E(y|\underline{x}) \leq 0.5 \\ 1 & if \quad 0.5 < E(y|\underline{x}) \leq 1.5 \\ 2 & if \quad 1.5 < E(y|\underline{x}) \end{cases} \tag{7}$$

Description length

- The cost of the model depends on the actual data, since we have to specify to the encoder the optimal predictions only for the seen prediction windows.
- Denoting n_x the number of different prediction windows found in the data set, the cost of encoding the model becomes $2n_x$ bits, and it is upper bounded by $2T$ where T is the number of measurements.

- The prediction errors will be encoded the same way as in the case of Boolean prediction. Therefore if N_e nonzero prediction errors are obtained with the Ternary predictor, the code length for the prediction residuals is $7N_e$.
- The overall coding length will be the model length plus the prediction error length:

$$L(n) = C_M(n) + 7N_e = 2n_x + 7N_e \text{ bits} \tag{8}$$

The perceptron

The perceptron is found by quantizing to three intervals the best linear combination of samples in the predictor window:

$$\hat{y} = \begin{cases} 0 & if \quad \underline{w}^T\underline{x} + w_0 \leq 0.5 \\ 1 & if \quad 0.5 < \underline{w}^T\underline{x} + w_0 \leq 1.5 \\ 2 & if \quad 1.5 < \underline{w}^T\underline{x} + w_0 \end{cases} \tag{9}$$

The parameters $\underline{w}$ and w_0 can be found using the Perceptron algorithm.

Description length

For ternary valued data, the numbers of all distinct perceptrons with two or three inputs are known. Using a perceptron model with two input genes requires

$$n_{model} = \log_2 471 = 8.88b \tag{10}$$

A perceptron model with three input genes requires

$$n_{model} = \log_2 85629 = 16.38b \tag{11}$$

A comparison of the three classes of models in a “12 gene” experiment

Hard Boolean Predictor			
	AHA=HBP(RCH1,p53,PC-1)	AHA=HBP(RCH1,p53)	AHA=HBP(p53)
N_e	2	2	12
Length	17	$\boxed{16}$	85

Hard Ternary Predictor			
	AHA=HBP(RCH1,p53,PC-1)	AHA=HBP(RCH1,p53)	AHA=HBP(p53)
N_e	1	2	7
Length	23	$\boxed{22}$	53

Perceptron Predictor			
	AHA=HBP(RCH1,p53,PC-1)	AHA=HBP(RCH1,p53)	AHA=HBP(p53)
N_e	2	2	7
Length	30.38	$\boxed{22.87}$	53.24
c_d (MSE)[?]	0.946	0.785	0.624

"12 Gene" Experiment Description length for three predictors : Hard Boolean Predictors, Hard Ternary Predictor and Perceptron Predictor

Length	g_1	g_2	Method	MI
16.00	12	1	2	-0.095037
36.00	12	2	1	0.005804
43.00	12	5	1	-0.077831

"12 Gene" Experiment All interesting predictors with window size =2. Hard Boolean prediction is "Method 2" and Ternary prediction is "Method 1".

Length	g_1	g_2	g_3	Method	MI
3.00	12	5	1	2	0.062436
17.00	12	2	1	2	0.053712
17.00	12	3	1	2	0.038086
17.00	12	4	1	2	0.071282
17.00	12	6	1	2	0.025156
17.00	12	7	1	2	0.023245
17.00	12	8	1	2	0.025823
17.00	12	9	1	2	0.000770
17.00	12	10	1	2	0.180767
17.00	12	11	1	2	0.019636
28.00	12	5	2	1	0.016835
37.00	12	5	4	1	0.048582
38.00	12	6	5	2	0.090102
38.00	12	8	5	2	-0.007450
38.00	12	9	5	2	0.000366
40.00	12	4	2	1	0.013190
40.00	12	6	2	1	0.073501
40.00	12	7	2	1	-0.003639
40.00	12	8	2	1	0.015921
40.00	12	9	2	1	0.003329
40.00	12	10	2	1	0.053081
40.00	12	11	2	1	-0.005184
42.00	12	7	5	1	0.038228
44.00	12	3	2	1	0.012155
45.00	12	5	3	2	0.017648
45.00	12	10	5	2	0.063625
47.00	12	9	6	1	0.046933
47.00	12	11	5	1	-0.046061

“12 Gene” Experiment All interesting predictors with window size =3.

An artificial example

We use first an artificial example, where, by knowing the true state of the nature, we will be able to check the success of our procedure.

We assume to have 30 measurements, at times $i = 1, \ldots, 30$, of 40 variables (factors) $x_1(i), \ldots, x_{40}(i)$ and 1 target $y(i)$, and suppose the measurements are quantized to three levels: $0, 1, 2$. The number of sequences of 30 symbols with ternary values is $3^{30} = 2 \cdot 10^{14}$, exceeding hugely the 40 sequences we have in the experiment. Suppose there is one function $y(i) = f^*(x_{t_1}(i), x_{t_2}(i), x_{t_3}(i), x_{t_4}(i))$ which exactly describes the target, where $f^* : \{0, 1, 2\}^4 \rightarrow \{0, 1, 2\}$.

The data was generated by randomly choosing 40 sequences of 30 measurements, from all possible $2 \cdot 10^{14}$ ternary sequences, with an uniform prior on all sequences. The target function $f^* : \{0, 1, 2\}^4 \rightarrow \{0, 1, 2\}$ was selected in the following way: for any input window $(j_1, j_2, j_3, j_4) \in \{0, 1, 2\}^4$, the value $f^*(j_1, j_2, j_3, j_4)$ was sampled from the random variable with values $0, 1, 2$ and mass $p = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$. Then the target gene was constructed as the exact function $y(i) = f^*(x_1(i), x_2(i), x_3(i), x_4(i))$. Therefore the true window for the target is $[x_1(i), x_2(i), x_3(i), x_4(i)]$, referred for short as $[1, 2, 3, 4]$. In the following we show that Boolean and Ternary predictors can be used in conjunction with the minimum description length to select prediction window candidates.

Target alone The complexity of the target (no conditioning on other genes) is evaluated by three methods: (a) a priori uniform distribution on all sequencies, giving $L(y) = -\lceil \log_2 3^{30} \rceil = 48$ bits; (b) adaptive arithmetic coding, giving the value $L(y) = 47$ bits; (c) Move to Front coding (which favors correlated sequence), giving the length $L(y) = 52$ bits.

Since the target was generated as an uncorrelated sequence of ternary values, Move to Front gives rather different results than the first two methods, but we can safely assume that the complexity of the target is in the range 47–52 bits.

Window size=2 No predictor using only two genes can achieve a description length lower than the complexity of the target. In the left Table we list the best models of order two, those giving a descriptive length less than 72, which makes 17 windows in total. The "correct" window (which is of order four, and contains gene$_1$, gene$_2$, gene$_3$, and gene$_4$) has a "trace" in the list of the best models, occupying the second position with the window (gene$_1$, gene$_3$).

Window size=3 The predictors using three genes becomes more successful in describing the target. In the middle Table we list the best models of order three, those giving a descriptive length less than 55, which makes 28 windows in total.

Window size=4 We list the best models of order four (those having length less than 49, which makes 23 windows in total). In bold we show the "true" model, which was found in the 19'th position, ranked out of 91390 possible prediction windows. Observe that variations of the true window shown up frequently as winners in the final list.

Policies for the limitation of the search space

Limiting the search space without sacrificing at all the optimality of the determinations should be done considering possible exclusion situations, like the following:

- Suppose we found the optimal model $\hat{g}_i = \hat{g}_i(g_j, g_k)$, and the determination $L(\hat{g}_i|g_j, g_k)$ is smaller than the three gene model cost, i.e. $L(\hat{g}_i|g_j, g_k) < L_M(3)$. Then it is clear that considering the model $\hat{g}_i = \hat{g}_i(g_j, g_k, g_\ell)$, will not lead to the average length $L(\hat{g}_i|g_j, g_k, g_\ell)$ smaler than $L(\hat{g}_i|g_j, g_k)$.

$Descr.Length$	g_1	g_2
60.00	15	40
65.00	**1**	**3**
67.00	4	13
67.00	9	40
67.00	10	11
67.00	10	34
67.00	14	20
67.00	20	31
67.00	20	32
67.00	20	40
67.00	26	34
67.00	27	40
67.00	38	40
72.00	3	34
72.00	11	34
72.00	27	32
72.00	32	36

$Descr.Length$	g_1	g_2	g_3
44.00	1	3	12
47.00	10	32	36
48.00	16	20	31
48.00	30	32	38
49.00	12	15	40
49.00	14	19	23
50.00	2	10	32
50.00	10	31	34
50.00	15	20	32
50.00	22	33	34
52.00	6	20	25
52.00	8	10	12
52.00	8	20	32
52.00	13	18	19
53.00	**1**	**2**	**3**
53.00	10	11	34
53.00	10	15	40

$Descr.Length$	g_1	g_2	g_3	g_4
46.00	1	2	3	12
46.00	1	3	12	38
46.00	12	30	32	38
46.00	30	32	34	38
46.00	8	10	31	34
46.00	1	12	32	38
46.00	12	30	32	38
48.00	1	3	12	18
48.00	10	27	32	36
48.00	4	16	20	31
48.00	2	10	25	32
48.00	2	10	29	32
48.00	10	25	31	34
48.00	15	20	25	32
48.00	15	20	32	40
48.00	6	20	25	40
48.00	8	20	32	34
48.00	**1**	**2**	**3**	**4**
48.00	10	11	13	34
48.00	15	20	32	40
49.00	1	3	12	34
49.00	12	16	20	31

Artificial data The prediction windows of the best predictors of lengths two, three and four

A real microarray example: Predicting a target gene using several hundred predictor genes

We have applied the same techniques to a real data set, containing the expressions of 576 genes for 31 patients. We concentrate on predicting the values of three target genes, knowing the values of the other genes.

1 Target 3 (the 44'th gene in the data set)

Target alone One can evaluate the selfinformation of the target in several ways. By using adaptive arithmetic coding with zero order contexts, L(target(3))=35 bits. By using MoveToFront L(target(3))=40 bits, and by uniform enumeration of all possible sequences of 31 ternary values we get L(target(3))=49 bits.

Window size=2 There are quite many good predictors of order 2. In Figure 1 we show the description length obtained by all possible predictors with two inputs. We observe that about 1300 predictors give a description length better than that of the target alone.

The absolute best prediction windows give a description length of 18 bits (there are 8 of them).

We list in the left part of Table 1 the best prediction windows with two genes. We observe

441 appears 3 times, 546 and 538 appears 2 times.

$target(gene(44))$	1201000001201100000000000000000
$\hat{y}(gene(441), gene(546))$	0200000001201100000000000000000
$gene(441)$	2021222120021122222222212222222
$gene(546)$	1201021111222200010102111101100
$gene(538)$	1011111210010112111121222222222

To keep the computations at an acceptable level at the next stage, we keep only the best 339 prediction windows (those with description length less than 32), and evaluate the performance of the three gene prediction windows.

Window size=3 For three gene prediction windows we have kept the best 339 prediction windows with two inputs, and enlarged each of those in turn with all possible prediction genes (587 of them) to get 198 654 prediction windows with three genes.. In this way we obtain extremely many good predictors with three genes. In Figure 2 we show the description length obtained by the 198 654 predictors with three gene windows. We observe that about half of them give a description length better than that of the target alone.

The absolute best prediction windows give a description length of 15 bits (there are 101 of them). Several of them are listed in Table 7, middle.

Window size=4 For four gene prediction windows we have kept the best 403 prediction windows with three inputs, and enlarged each of those in turn with all possible prediction genes

(587 of them) to get 236 158 prediction windows with four genes. In this way we obtain extremely many good predictors with four genes. In Figure 2 we show the description length obtained by the 236 158 predictors with four gene windows. We observe that about $\frac{4}{5}$ of them give a description length better than that of the target alone.

The absolute best prediction windows give a description length of 16 bits (there are 235 of them). Several of them are listed in Table 7, right. We show in Figure 4 the histogram of how many time a window of three genes was extended successfully to a prediction window with four gene which gives description length 16. Most of the time the winners in the four gene case were extensions of very good predictors of order three. It is however possible for three gene predictors with a fair performance (the one we ranked as 400's) to be extended successfully, to reach the top place in the four gene case.

Applying strictly the MDL principle, we should prefer as models the prediction window with the smallest description length. Therefore $(21, 136, 253)$ is a more likely model than $(21, 136, 253, 176)$ (including the gene 176 as a factor does not pay off, in terms of descriptive power of the model).

But since our goal is to just signal the best candidates, we will signal both $(21, 136, 253)$ and $(21, 136, 253, 176)$ windows, since both have remarkable low description length, when compared to other predictor combinations.

2 Target gene 1

We dedicate a special section to target gene 1, which has the following values along the 31 measurements $-1, 0, -1, -1, -1, -1, -1, \ldots, -1$. It is obvious that the experiment is not well designed with respect to this target gene, it brings almost no information about the relation of this gene with the rest of the genes.

One can evaluate the selfinformation of the target in several ways. By using adaptive arithmetic coding with zero order contexts, L(target(1))=14 bits. By using MoveToFront L(target(1))=47 bits, very close to the uniform evaluation of all possible sequencies of 31 ternary values which is L(target(1))=49 bits.

We found 366 prediction models with two genes who were able to model perfectly the target gene 1 (prediction residuals are 0).

All combinations of two gene in the prediction window were able to achieve description lengths lower than the selfinformation of the target sequence (there are 171092 such combinations), the maximum length in this experiment being $\max L(2) = 11$ bits. Compare the worst case (11 bits) to the 14 bits which is the description length of the target without any conditioning.

One example of perfect model is the prediction window $[gene(331), gene(586)]$, and the Boolean predictor $f(x_1, x_2) = \overline{x}_1 x_2$. The Boolean prediction has a very clear meaning: $gene(20)$ should be 1 whenever $gene(331) < gene(586)$. The 31 values of each gene and prediction are shown in the following table:

$gene(331)$	2021112211121121111211112211111
$gene(586)$	1100011101000000010000001011000
$\hat{y}(gene(331), gene(586))$	0100000000000000000000000000000
$gene(20)$	0100000000000000000000000000000

The conclusion to draw for target gene 1: (a) the experimental data is not informative, since any possible combination of two factors will describe well the data (b) if one wants to extract some information out of this experiment, (with very uninformative data), a ranking of individual factors can be done by studying the "frequency of determination"(how many times a factor appeared in prediction windows having small description lengths).

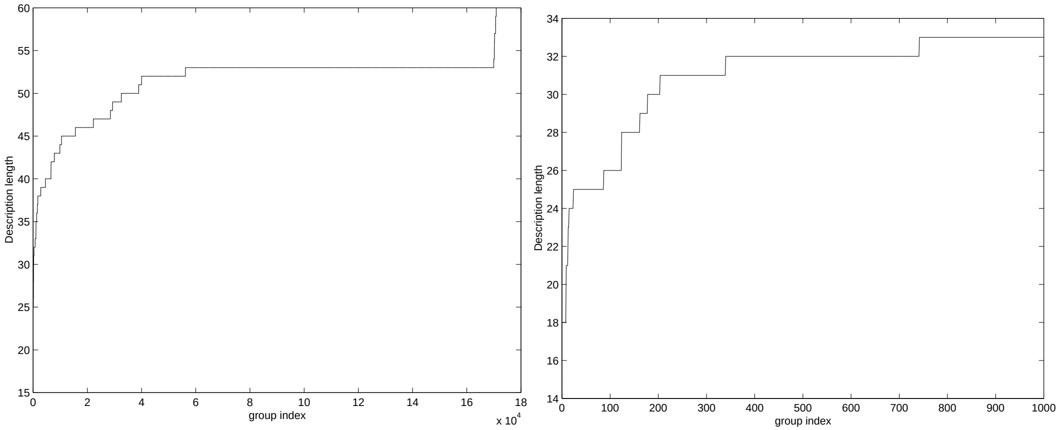

Figure 1: **587 Gene Experiment** The description length obtained by all predictors with two inputs, for the target gene 3 (44'th in prediction set). The best 339 windows were selected to be extended to 198 654 three–gene windows.

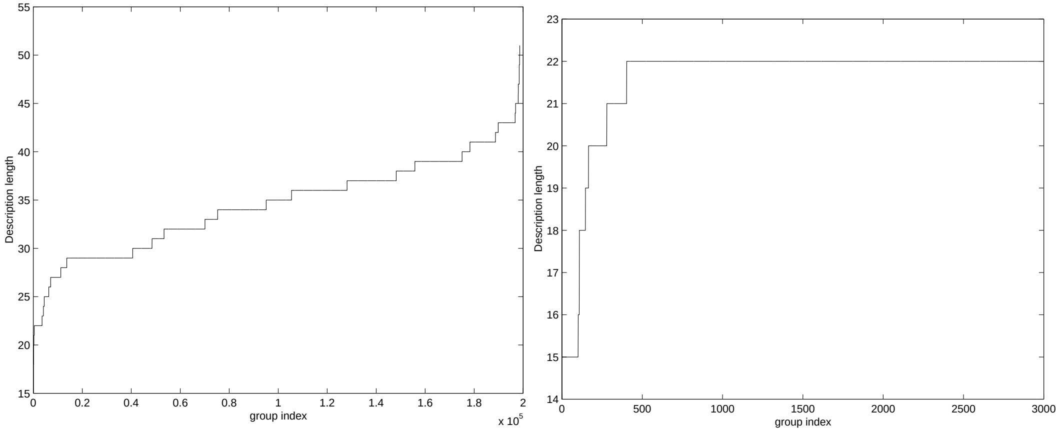

Figure 2: **587 Gene Experiment** The description length obtained by predictors with three inputs, for the target gene 3 (we tested only 198 654 out of 33 538 245 prediction windows). The best 403 windows were selected to be extended to 236 158 four–gene windows.

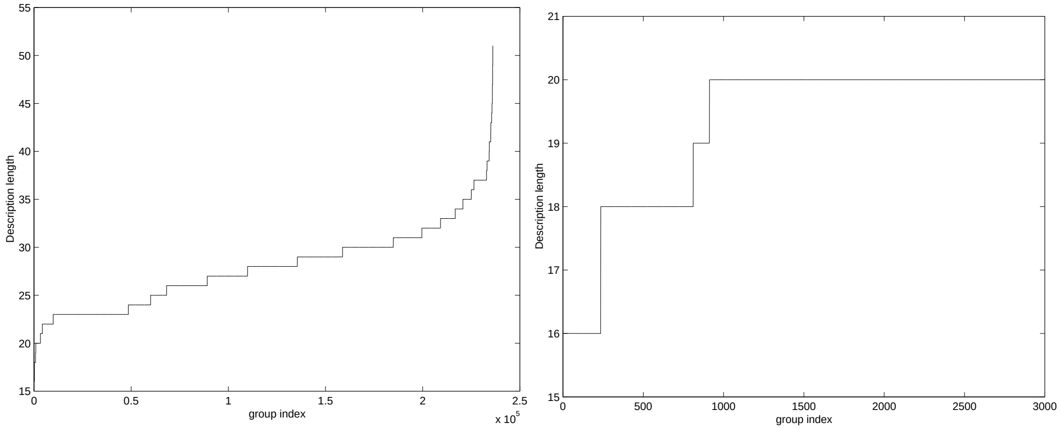

Figure 3: **587 Gene Experiment** The description length obtained by predictors with four inputs, for the target gene 3 (we tested only 236 158 out of 489 658 377 prediction windows)

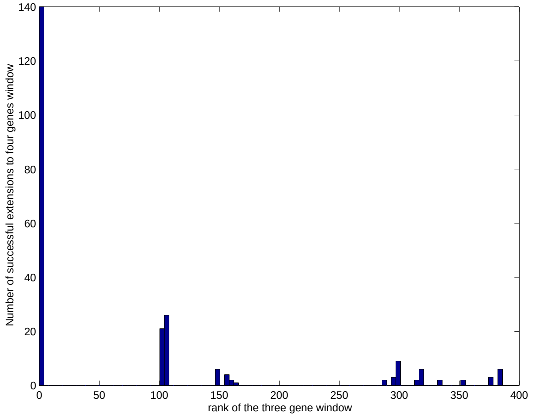

Figure 4: **587 Gene Experiment** The histogram of how many times a window of three genes was extended successfully to one of the best prediction windows with four genes (those which give description length 16).

Descr.Length	g_1	g_2
18.00	21	136
18.00	38	538
18.00	53	471
18.00	106	204
18.00	320	441
18.00	441	538
18.00	441	546
18.00	503	546
21.00	53	398
21.00	53	501
21.00	136	298
21.00	204	501
23.00	186	398
23.00	398	538
24.00	106	441
24.00	136	146
24.00	136	155
24.00	136	250
24.00	136	281
24.00	136	354
24.00	152	441
24.00	441	466
24.00	441	585

Descr.Length	g_1	g_2	g_3
15.00	21	136	81
15.00	21	136	91
15.00	21	136	104
15.00	21	136	110
15.00	21	136	158
15.00	21	136	238
15.00	21	136	250
15.00	21	136	253
15.00	21	136	262
15.00	21	136	279
15.00	21	136	298
15.00	21	136	308
15.00	21	136	332
15.00	21	136	335
15.00	21	136	359
15.00	21	136	408
15.00	21	136	442
15.00	21	136	476
15.00	21	136	504
15.00	21	136	522
15.00	21	136	523
15.00	21	136	538
15.00	21	136	548
15.00	21	136	578
15.00	38	538	288
15.00	106	204	176
15.00	441	538	288
15.00	441	546	314
15.00	503	546	21
15.00	503	546	31
15.00	503	546	300
15.00	503	546	389
15.00	503	546	468

Descr.Length	g_1	g_2	g_3	g_4
16.00	21	136	253	176
16.00	21	136	253	263
16.00	21	136	253	467
16.00	21	136	253	550
16.00	21	136	262	412
16.00	21	136	262	476
16.00	21	136	262	568
16.00	21	136	279	119
16.00	21	136	279	176
16.00	21	136	279	178
16.00	21	136	279	243
16.00	21	136	279	260
16.00	21	136	279	321
16.00	21	136	279	432
16.00	21	136	279	445
16.00	21	136	279	530
16.00	21	136	279	545
16.00	21	136	279	552
16.00	21	136	279	565
16.00	21	136	308	568
16.00	21	136	335	176
16.00	21	136	476	34
16.00	21	136	476	41
16.00	21	136	476	46
16.00	21	136	476	67
16.00	21	136	476	69
16.00	21	136	476	126
16.00	21	136	476	142
16.00	21	136	476	146
16.00	21	136	476	153
16.00	21	136	476	155
16.00	21	136	476	176

Table 1: **587 Gene Experiment** Target gene 44: The prediction windows of the best predictors of lengths two, three and four