

ActiveSeqAnalyzerTN のマニュアル

Introduction

このプログラムは、生物活性と塩基配列を関連付け、その生物活性に関連した最適塩基配列を抽出する解析プログラムです。表 1 のようなデータから、活性値に対するポジション別の各塩基の重みを計算することで、活性値を最も増加（減少）させる配列を抽出します。

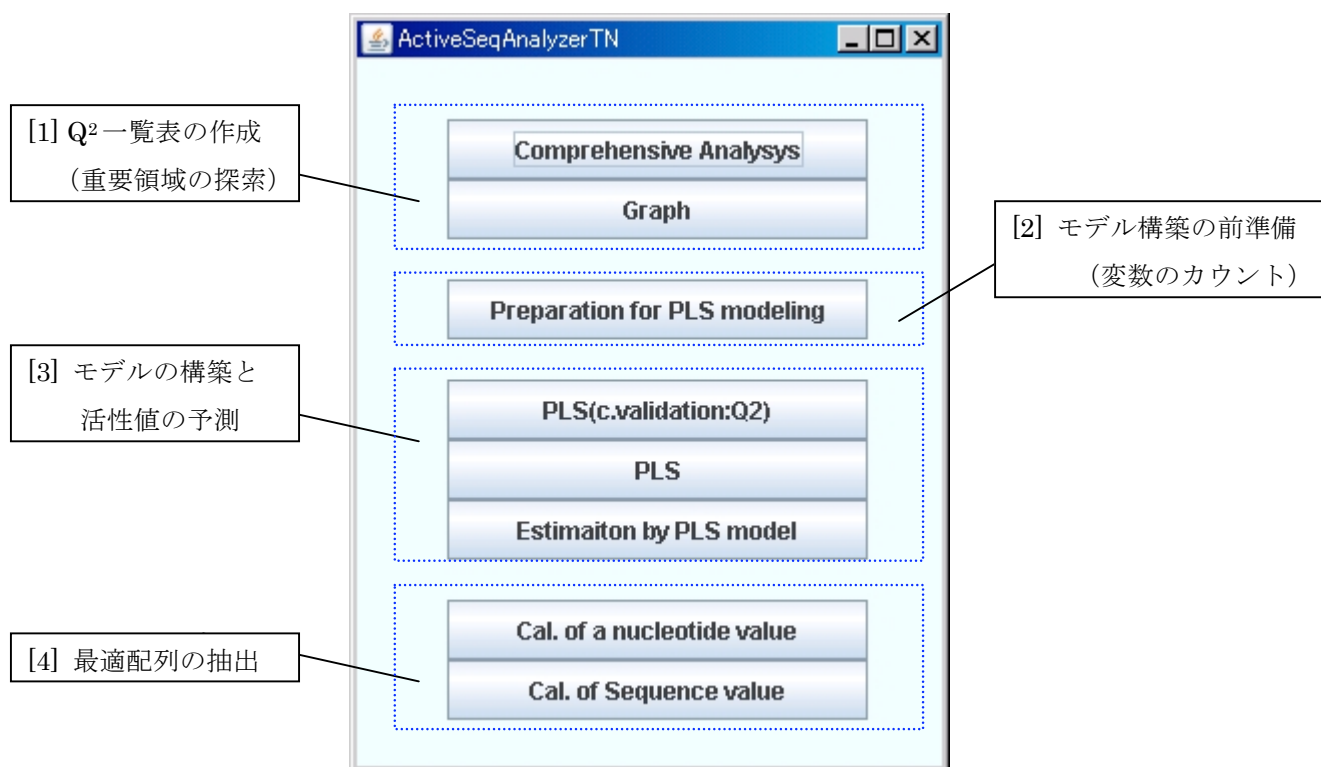


図 1 メイン画面

【解析の流れ】

重要領域の探索を行い([1])、その結果を参考に PLS モデルを構築し([2],[3])、得られたファイルを使用して最適配列を抽出する([4])。

[1]-1 【Comprehensive Analysis】： Q²一覧表の作成（重要領域の探索）

入力ファイル名：Seq* 出力ファイル名：ComQ2*

重要領域を探索するための Q²一覧表を作成します。

下記設定項目にて設定した領域について網羅的にモデルを構築し、その Q² 値の一覧を”ComQ2 ファイル”に出力します。

- ① 読み込むファイルを選択する。

- ② 各種項目を設定する。
- ③ 【Comprehensive Analysis】をクリックして解析を開始します。
※長い計算時間を要することがあります。

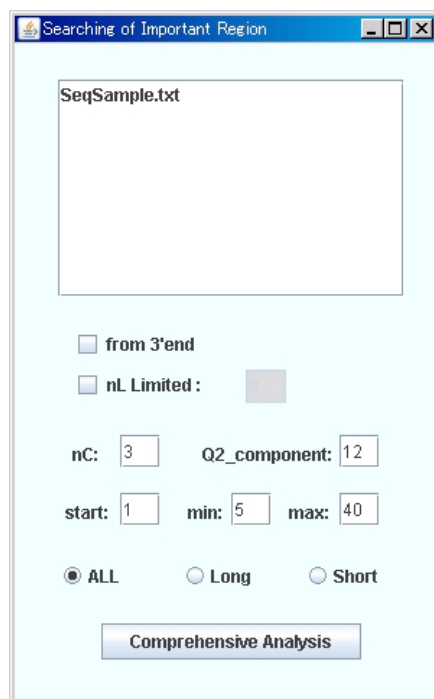


図 2 設定画面 (Comprehensive Analysis)

《設定項目》

・ from 3'end(チェックボックス):

チェックを入れることで 3'端(開始コドン側)を基点とした解析をおこないます。

・ nL Limited (チェックボックス) :

チェックを入れ、任意の塩基長を設定します。これにより任意の塩基長のモデルのみを解析対象とします。

・ nC :

解析する部分塩基配列の長さ

nC=3 のとき、"nnn"という 3 塩基配列をカウントします。

・ Q²_component:

PLS を行うときの潜在変数の数

配列データから抽出した変数の数が潜在変数の数より小さい場合エラーが出ます。

• start:

調べる領域の開始位置

• min:

調べる領域の最小の位置

nL Limited にチェックを入れた場合は不要です。

• max:

調べる領域の最大の位置

例)

start = 1, min = 5, max = 40 の場合

最小の領域が 1~5, 2~6 , 3~7 ...

最大の領域が 1~40

となり、1~40 までに存在するすべての領域を網羅的に解析します。

• ALL: すべてのサンプルを解析の対象とします。

• Long: 長さが max 以上のサンプルのみを解析の対象とします。

• Short: 長さが max 以下のサンプルのみを解析の対象とします。

[1]-2 【Graph】

入力ファイル名 : ComQ2*

Comprehensive Analysis の結果をグラフに表示します。これにより重要領域の推測を行います。

① 表示したいファイルを選択

② 【Graph】をクリックして、グラフを表示します。

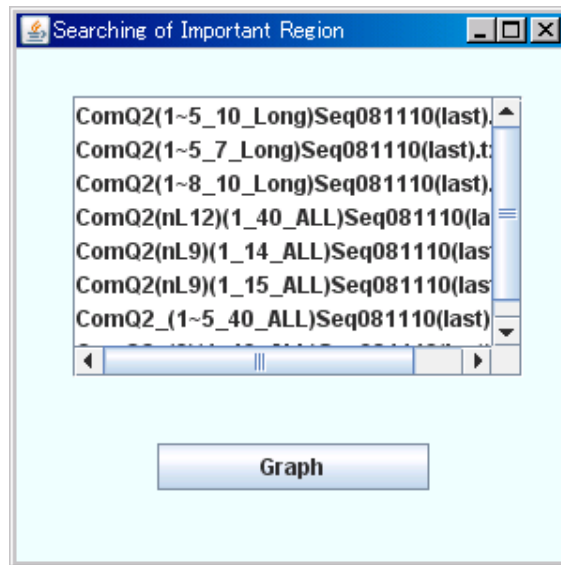


図 3 グラフデータの選択画面

《グラフの見方》

横軸が塩基のポジション、縦軸が Q^2 を表し、1つの横線が1つのデータ(領域と Q^2)を表しています。 Q^2 の高い領域が重要領域として考えられます。

・メニューの説明

[File]-[Open] : ComQ2 ファイルを開きます。

[File]-[Save] : 現在表示されているグラフを bmp 形式で保存します。

[File]-[Exit] : グラフ画面を閉じます。

[Setting]-[from 3'end] : 3'端を基点としてグラフを表示します。

※ 3'端側を基点とした場合、データは左右反転して表示され、画面右側から左側にかけて、塩基位置が増加します。(3'端側から解析したデータを表示する場合に使用します。)

[Setting]-[# of Samples] : 解析に使用したサンプル数を表示します。

※ データ中のサンプルの塩基長が異なると、解析に使用するサンプル数も長さにより異なってきます。(サンプル数の影響を考慮する場合はチェックを入れます。)

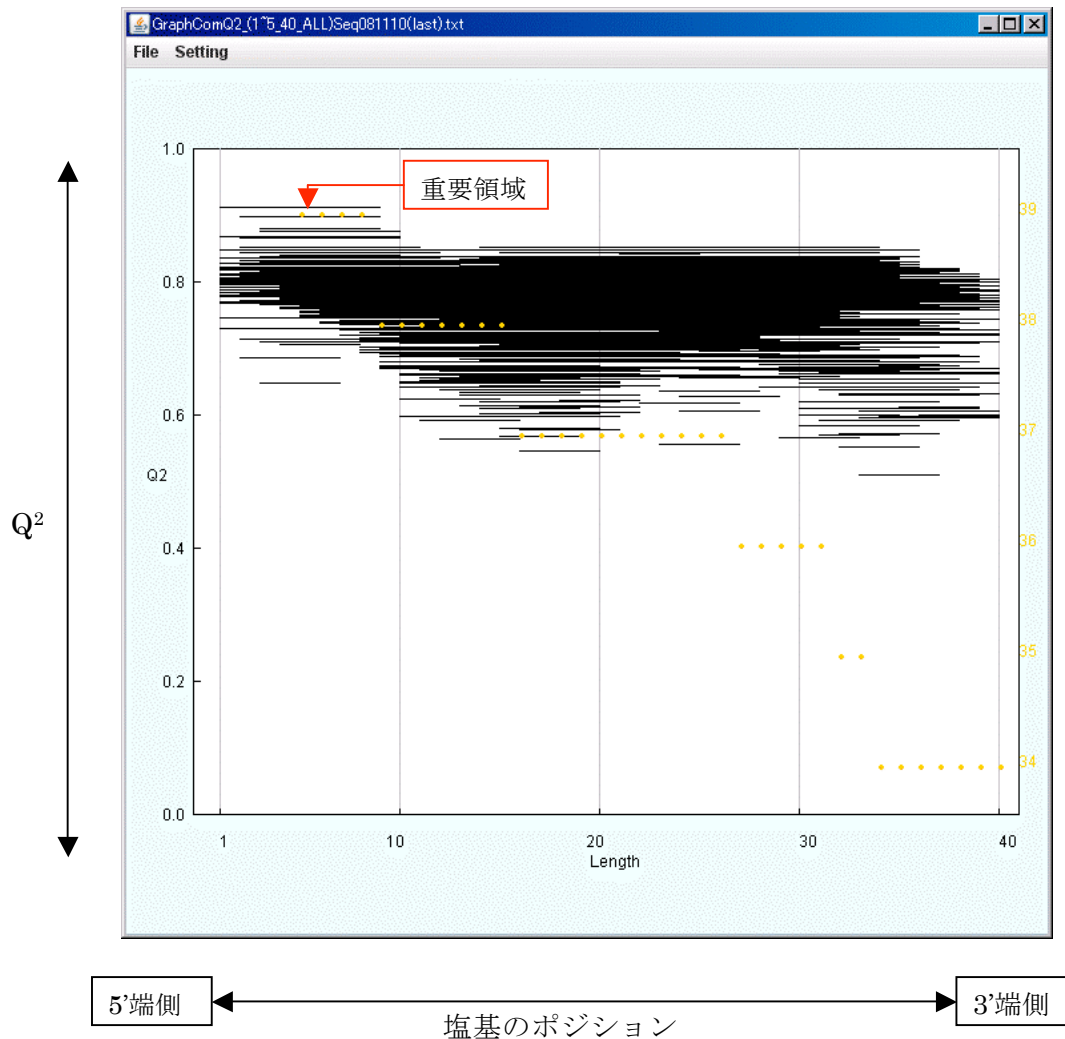


図 4 5'端を基点としたグラフ

Q^2 の高い領域を重要領域として考えるので、この図の場合は1~9が重要領域であると推測します。黄色のプロットおよび文字がサンプル数を示しています。

[2] 【Preparation for PLS modeling】モデル構築の前準備（変数のカウント）

入力ファイル名：Seq* 出力ファイル名：PLSData*

PLS モデルの構築のための前準備をします。配列データから変数をカウントし結果を出力します。

- ① 下記設定項目を設定します。
- ② 【Count】をクリックし解析を開始します。

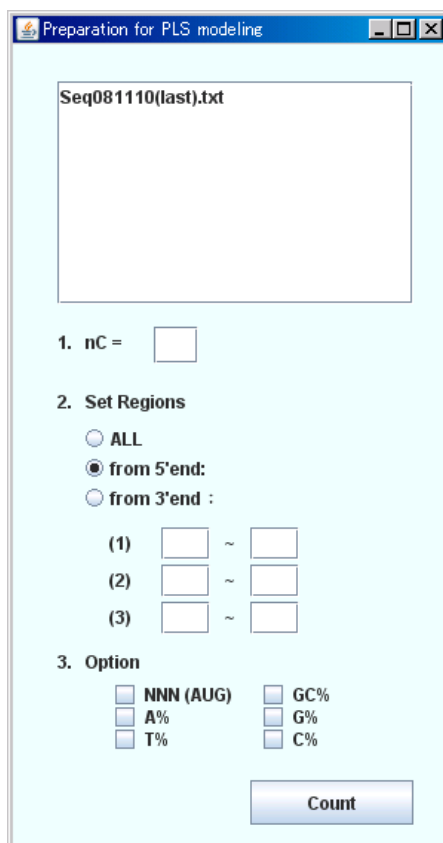


図 5 Preparation for PLS modeling(設定画面)

《設定項目》

1. nC：部分塩基配列の長さ

2. Set Regions (変数をカウントする範囲を指定します。)

ALL: それぞれのサンプルの全長を対象にします

from 5\'end: 5\'端から任意の長さの範囲を対象にします。

from Start codon: 3\'端から任意の長さの範囲を対象にします。

Multiple Regions: 任意の範囲を対象にします

最大で 3 つの範囲を対象とした複合予測モデルを構築
 できます。このとき、それぞれの範囲でカウントした
 変数はそれぞれ独立したものとして扱われます。

from 5\'end 5\'端側から任意の範囲を対象にします。

from Start codon 3\'端側から任意の範囲を対象にします。

PLS 回帰モデルを構築します。

- ① PLS の軸数を設定する。(カウントした変数の数より低いと計算できません)
- ② "PLSData"ファイルを選択します。
- ③ **【Start】** をクリックして計算を開始します。

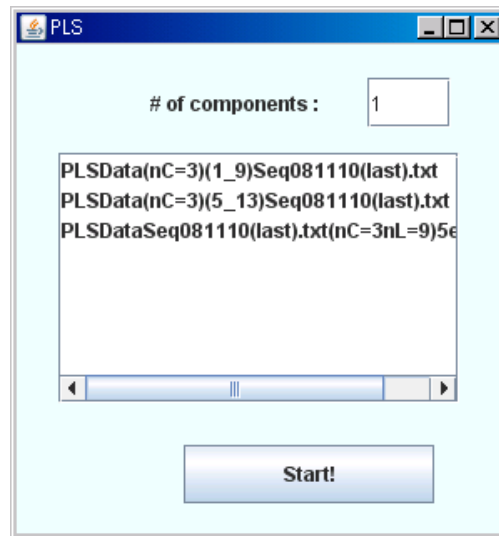


図 7 PLS (ファイル選択画面)

[3]-3 【Estimation by PLS model】 ※最適配列の抽出に関してこのステップは不要です。

入力ファイル名 : PLSData* 出力ファイル名 : CoefPLS*, PLS(n)*

任意の"CoefPLS"ファイルを使い、任意の配列データ("PLSData"ファイル) についてその活性値を予測します。

- ① 使用する"CoefPLS"ファイルを選択します。
- ② 解析する配列データを選択します。(事前に"Seq"ファイルから"PLSData"ファイルを出力しておく必要があります。)
- ③ **【Start】** をクリックして計算を開始します。

※ "CoefPLS"ファイルに存在しない変数をもつ配列データは除外されます。
除外せずに (PLS 係数が存在しない変数については係数を 0 として扱う) 全てのデータを解析する場合は、チェックボックス(ALL)にチェックしてください。

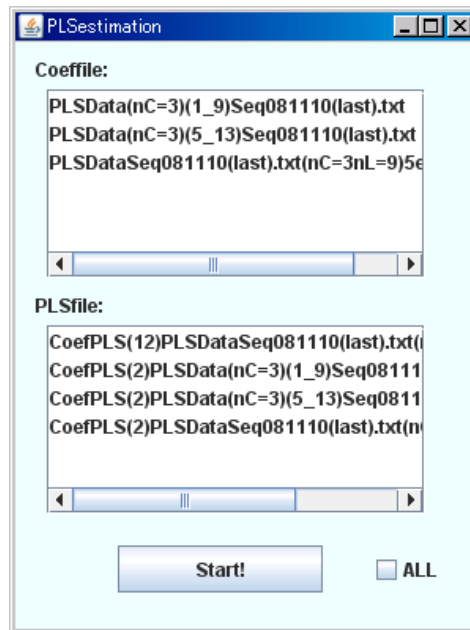


図 8 Estimation by PLS (ファイル選択画面)

[4]-1 【Cal. Of a nucleotide value】:最適配列の抽出 (塩基の重み付け)

入力ファイル名 : CoefPLS*,Seq* 出力ファイル名 : NtValue*

ポジション毎の各塩基の重み付けを行います。

- ① 基点とする側と、解析する領域を設定します。(使用する係数ファイルに合わせて下さい)
- ② 使用する係数ファイル([3]で出力されたもの)を選択します。
- ③ 解析する配列データを選択します。
- ④ **【Start】** をクリックして解析を開始します。統計的に有意かつ最も高い値を示す塩基を並べた配列が最適配列です(有意なものがない場合や、正のものがない場合はもっとも高い値を示す塩基を選択します)。なお、後述の**【Cal. Of Sequence value】** を実行することで、自動的に最適配列を出力させることができます。

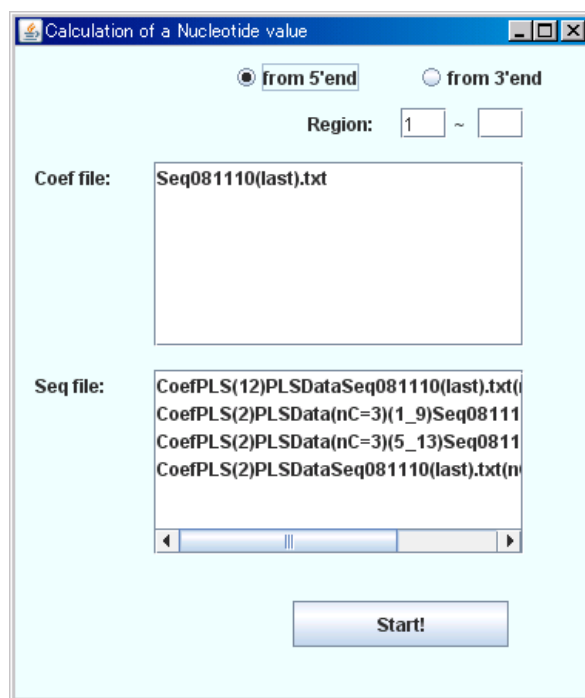


図 9 Calculation Of a nucleotide value (ファイル選択画面)

[4]-1 【Cal. Of Sequence value】:最適配列の抽出 (塩基の重み付け)

入力ファイル名 : CoefPLS*,Seq* 出力ファイル名 : Svalue*

”NtValue”ファイルをもとに、最適配列の抽出と、各配列の重み付けを行います。
 ※”NtValue”ファイルを参照し、各塩基の重みの総和をとったものを配列の重みとして
 しています。

- ① ”NtValue”ファイルと、”Seq”ファイルを選択します。
- ② P-value を設定し閾値を決めます。
- ③ 【Start】 をクリックして計算を開始し、結果を Svalue ファイルに出力しま
 す (この際、最適配列の出力と同時に、”Seq”ファイル中の各配列についても
 重み付けを行います。なお配列の重み付けに関して、統計的に有意でない塩
 基については、その重みを 0 として計算しています)。

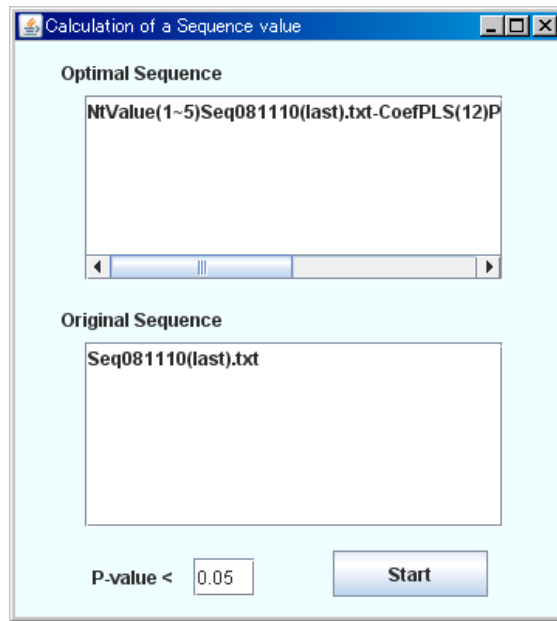


図 10 Calculation Of Sequence value (ファイル選択画面)

各種ファイルの説明

Seq**

解析の元となるデータです。

Sample_name	Value	Sequence
Sample 1	8.4748	aatcctgtggggattttttttttt
Sample 2	3.9753	ggtcattgtttattag
⋮	⋮	⋮
Sample N	16.2322	attttaaatggccccgtttgtat

表 1 Seq ファイルのフォーマット

1 列目：サンプルの名前、2 列目：活性値、3 列目：塩基配列

1 行目は解析されませんが必要ですので、適当なラベルを入れておいてください。

ComQ2_**

Q² の網羅的解析で出力されるファイルです。またグラフで表示する際に読み込むファイルです。

Filename	Max_Q2	Components	# of sample
1~5	0.688996	1	39
1~6	0.663933	4	39
⋮	⋮	⋮	⋮
36~40	0.706775	3	34

表 2 ComQ2 ファイルのフォーマット

1 行目に読み込んだファイル名とラベルが出力されます。

1 列目：解析した領域、2 列目：最大 Q²、3 列目：最大 Q² を出すときの components の数、4 列目：解析の対象となったサンプル数

PLSData**

Seq ファイルから変数を抽出したファイルです。PLS 回帰モデルを構築するときに使用します。

>filename	Sample 1	Sample 2	Sample N
>Target	0.01	0.20	1.00
Variable 1	3	1	10
Variable 2	5	0	2
⋮	⋮	⋮	⋮	⋮
Variable M	2	7	0

表 3 PLSData ファイルのフォーマット

1 行目にサンプル名、2 行目に活性値、3 行目以降に各変数のパラメータが出力されます。

PLS(Q2CrossValidation)**

PLS 回帰において、軸数を変えたときのそれぞれの Q^2 と各サンプルに対しての予測値が出力されます。このファイルから最大 Q^2 を算出するときの潜在変数の数 (components) を調べます。

>PLS(Cross Validation)			
1	># of components=	1	
1	sampleID :0	8.4748	16.97266
1	sampleID :1	3.9735	10.94786
1	sampleID :N	16.2322	6.597798
1	Q2=	0.8200521	
2	># of components=	2	
2	sampleID :0	8.4748	15.63982
2	sampleID :1	3.9735	5.963098
2	sampleID :N	16.2322	10.35826
2	Q2=	0.8679568	
⋮	⋮	⋮	⋮
K	># of components=	K	
K	sampleID :0	8.4748	14.67989
K	sampleID :1	3.9735	8.799194
K	sampleID :N	16.2322	15.27969
K	Q2=	0.718372	

表 4 PLS(Q2CrossValidation)ファイルのフォーマット

赤枠で囲った部分を参照し、components の数を決定します。

CoefPLS(*)**

各変数の PLS 係数を出力したファイルです。()内は軸数です。[5]最適配列を抽出する際に使用します。ここで出力されている結果は、PLS(*)**ファイルにも出力されるので、そちらを参照してください。

PLS(*)**

構築したモデルについての情報が出力されます。ファイル名の()内は軸数です。

>PLS		
>Components =	10	
>Rpred2 =	0.99827	
Sample Name	y(Original)	Y(predicted)
Sample 1	8.4748	8.317337
Sample 2	3.9753	3.636633
⋮	⋮	⋮
Sample N	16.2322	16.69105
>Model Equation		
y(pred) =	4.070566	
Variable 1	0.403555	
Variable 2	0.738943	
⋮	⋮	
Variable M	1.882595	

実測値

予測値

PLS 係数

表 5 PLS(*)**ファイルのフォーマット

上段に選択した components の数、Rpred、各サンプルの活性値と予測値、下段に各変数の PLS 係数が出力されます。

Est**

任意の配列データについて、あらかじめ構築していたモデルを使用したときの予測値が出力されます。

Sample 1	8.317337
Sample 2	3.636633
⋮	⋮
Sample N	16.69105

表 6 Est**ファイルのフォーマット

1 列目：サンプル名、2 列目：オリジナルの活性値、3 列目：予測モデルによる予測値

NtValue**

各ポジションにおける塩基の、活性値に対する寄与度とその統計的な検定量(t 値と p 値)が出力されます。このファイルを参考に、統計的に有意で、寄与度が最も大きい塩基を各ポジションから抽出し、最適配列とします (このファイルを【Cal. Of Sequence value】

に供することで自動的に最適配列が出力されます)。統計的に有意な塩基がない場合は寄与度の最も大きい塩基を抽出します。また、ファイル下部には、各サンプルについて PLS 係数を配置したときの数値が出力されます。

>Pos.	1	2		x
a	0.01242	0.053982		0.213496
c	0.02053	0.020312	0.023274
g	-0.26528	-0.16343		-0.06176
t	0.203302	0.136019		-0.09417
} 各塩基の重み				
>t-value				
a	-0.64209	0.747744		4.725264
c	0.235961	0.268696	0.393336
g	-4.97388	-3.85075		-2.78461
t	1.295527	2.988645		-1.3822
} t 値				
>p-value				
a	0.26296	0.22967		0.000164
c	0.407382	0.394831	0.348164
g	0.000765	0.00026		0.004198
T	0.10158	0.002478		0.087595
} p 値				
Sample 1	g	c	c
	-0.19501	-0.01004		-0.05905
Sample 2	a	c	c
	0.001843	0.023211		-0.18296
⋮	⋮	⋮	⋮	⋮
Sample N	a	c	g
	0.168598	0.339402		-0.00311

表 7 NtValue ファイルのフォーマット

赤枠で囲った部分を参照し、最適配列を抽出します。

Svalue**

NtValue ファイルを基に、最適配列とその配列の重み、および任意の Seq ファイル中の配列について配列の重みが出力されます。この配列の重みを指標に、活性値に対する最適配列の寄与度とオリジナルの配列の寄与度を比較することができます。

SampleName	Original Value	Seq_Value	Sequence
Optimal_Sequence		16.30994	(5')1- (a)(a)(t)ta(a)(a)aa- 9
Sample 1	8.4748	3.554195	aatcctgtggggattttttttttt
Sample 2	3.9753	5.807796	ggtcatgtttattag
⋮	⋮	⋮	⋮
Sample N	16.2322	2.795914	attttaaatggccccgtgtttgtat

表 8 Svalue ファイルのフォーマット

赤枠で囲った部分が最適配列です。両端の数字はその配列の塩基位置を表しています。(この場合、5'端 1~9 の aattaaaa が最適配列です。3'端を基点としている場合は 9-aattaaaa-1(3')のように左右の数字が逆転し右端に(3')と表示されます。なお、()は統計的に有意なものが無かったことを意味します。)

