KaPPA-Average 1.0

マニュアル

第 1.0 版

制作者: かずさ DNA 研究所　櫻井　望

制作日: 2010 年 1 月 12 日

目次

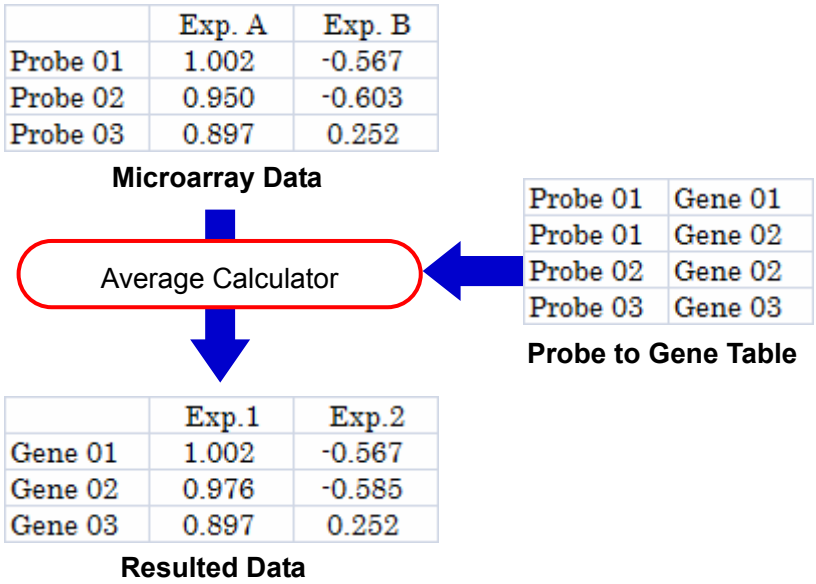
1. はじめに

1-1. KaPPA-Average とは

KaPPA-Average は、KaPPA-View（http://kpv.kazusa.or.jp/kpv4/）でマイクロアレイデータを解析する際に便利なデータ変換ソフトウェアです。

一般のマイクロアレイでは、一つのプローブが複数の遺伝子に対応していることがあるので、遺伝子を主体に考えた場合、どのプローブのデータを採用して良いか判断に迷うことがあります。KaPPA-View に搭載されている遺伝子情報も、例えばモデル植物シロイヌナズナでは、TAIR（http://www.arabidopsis.org/）が整備している AGI 番号で管理されているため、各社のマイクロアレイのプローブ番号との対応を考える必要があります。

KaPPA-Average では、プローブと遺伝子間の多対多の関係を考慮して、アレイで検出されたプローブごとのデータを、遺伝子ごとのデータに変換することができます。ひとつの遺伝子に複数のプローブが対応している場合、遺伝子のデータは、対応するプローブの平均として計算されます。



KaPPA-View は、遺伝子発現データを代謝マップへあてはめることにより、変動の傾向を大まかに理解することを主な目的としていますので、このような平均化処理は、

解析の最初のステップとしては有効と思われます。

1-2. 動作環境

KaPPA-Average は Java で作成されたソフトウェアです。ご使用には、Java Runtime Environment 1.5.0 以上がインストールされた OS が必要です。

OS:

Windows XP/Vista (Microsoft)

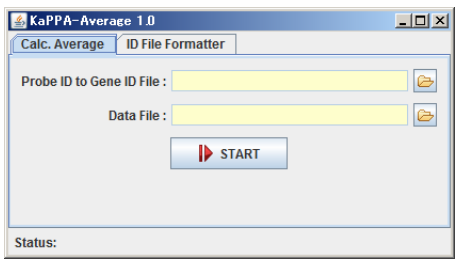
Mac OS X (Apple)

Linux

1-3. インストールと起動

ダウンロードした zip ファイルを解凍し、適当な場所に保存してください。解凍したフォルダに含まれる KaPPA-Average.jar をダブルクリックすると、ソフトウェアが起動します。

起動画面



2. 操作説明

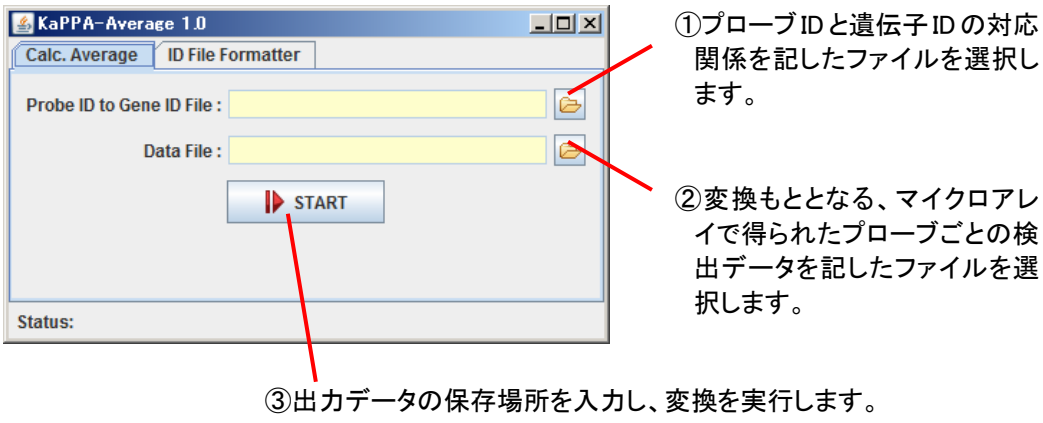
KaPPA-Average には、以下の二つの機能があります。

１．メインの機能
マイクロアレイで得られたプローブごとのデータから、プローブ ID と遺伝子 ID の対応表をもとに、遺伝子ごとのデータに変換します。

２．プローブ ID と遺伝子 ID の対応表を作成するための支援機能
タブ区切りテキストなどを処理して、１．で使用する対応表を作成します。

2-1. メイン機能 - Calc. Average

「Calc. Average」のタブでは、本ソフトウェアのメイン機能である、データ変換を行います。



2-1-1. データの準備

①ID 対応ファイル
以下のように、1 列目にプローブ ID、2 列目に遺伝子 ID が記述されたタブ区切りテキストをご用意ください。ヘッダー行はなく、1 行目からデータが始まっていることにご

注意ください。

	A	B
1	A_84_P10173	AT5G10880
2	A_84_P534845	AT2G01390
3	A_84_P712221	AT1G72860
4	A_84_P712221	AT1G72855
5	A_84_P15539	AT3G24350
6	A_84_P756685	AT2G14370
7	A_84_P820213	AT1G20560
8	A_84_P18685	AT5G11510
9	A_84_P19533	AT4G29060
10	A_84_P586827	AT5G08185
11	A_84_P21291	AT3G54300
12	A_84_P14903	AT5G11920
13	A_84_P539068	AT4G39290
14	A_84_P755862	AT2G07480
15	A_84_P175641	AT3G19660
16	A_84_P818748	AT3G13560
17	A_84_P727905	AT1G14650
18	A_84_P14847	AT4G40060

一つのプローブ ID が複数の遺伝子 ID に対応している場合などは、それぞれ個別の行として記載します（上図の 3 列目と 4 列目など)。

②データファイル

マイクロアレイで得られた、プローブごとの検出データは、以下のようなフォーマットとしてご準備ください。

	A	B	C
1		Exp. A	Exp. B
2	A_84_P10173	-0.287306	0.007288
3	A_84_P534845	-0.343136	0.0348
4	A_84_P712221	0.0927705	-0.01319
5	A_84_P15539	-0.104371	-0.00131
6	A_84_P756685	0.1512154	0.011672
7	A_84_P820213	0.2254475	-0.00256
8	A_84_P18685	-0.306554	0.013891
9	A_84_P19533	-0.094507	0.007606
10	A_84_P586827	0.1035178	-0.00398
11	A_84_P21291	0.1034709	-0.02099
12	A_84_P14903	0.1637749	0.025303
13	A_84_P539068	0.1184364	0.004917
14	A_84_P755862	-0.072946	0.001363
15	A_84_P175641	-0.529631	-0.02004
16	A_84_P818748	0.3484821	-0.00188
17	A_84_P727905	-1.311527	0.021977
18	A_84_P14847	-0.181841	-0.00804

1 行目：　ヘッダー行（必須）

2 列目以降のデータに対する実験名（データ名）を記入してください。

2 行目以降：　データ部分（必須）

1 列目にプローブ ID、2 列目以降に各実験で得られたデータ（数値）を記入してください。

※実験データ（2 列目以降）は何列あってもかまいません。

ファイル形式：　タブ区切りテキストとして保存してください。

注意

・データ部分に空白や数値以外の文字が入力されていると、エラーとなります。
・プローブ ID が重複しないようにしてください。重複していた場合、一番下の行に書かれたデータのみが有効になります。

2-1-2. 実行

２つのデータファイルが選択された状態で「START」ボタンを押すと、変換後のデータを保存するファイル名を問い合わせるダイアログボックスが開きます。保存するファイル名を入力すると、処理が始まります。
Status に「Finished.」という文字が現れれば、処理は終了です。

2-1-3. 出力データ

変換後の出力ファイルは、以下のようなタブ区切りテキストファイルとなっています。

	A	B	C
1		Exp. A	Exp. B
2	AT3G04220	0.191343591	−0.03366746
3	AT2G38110	−0.601452829	0.002557445
4	AT5G11420	−0.280091644	0.003028008
5	AT3G16650	0.130034958	−4.46E−04
6	AT5G02560	−0.098181204	−1.72E−04
7	AT2G31200	0.081133448	−0.004042342
8	AT4G26950	0.20754731	3.68E−04
9	AT5G25350	−0.175170294	0.00492307
10	AT2G27080	−0.1222054	−0.00331908
11	AT4G08800	−0.181022991	−0.002256198
12	AT3G10950	−0.124063125	−2.91E−04
13	AT3G26020	−0.260583756	0.001995765
14	AT1G47420	0.185523149	0.00187916
15	AT1G53270	−0.26798637	−0.013514267
16	AT2G17760	−0.323237129	−0.014847535
17	AT3G55665	−0.175394895	0.003699085
18	AT5G09810	0.795378655	0.006704669

また、出力ファイル名の拡張子が「.log」となっているファイルも同時に出力されます。
（例えば、出力ファイル名が「result.txt」であれば、「result.log」）

	A	B	C	D
1	Probe ID	Gene ID	Exp. A	Exp. B
2	Average	AT1G29970	0.165854093	0.003318564
3	A_84_P55970	AT1G29970	-0.03611878	0.006839025
4	A_84_P808527	AT1G29970	0.274866829	0.01994654
5	A_84_P599601	AT1G29970	0.293861297	-0.001138482
6	A_84_P867964	AT1G29970	-0.173973874	-0.002314763
7	A_84_P808508	AT1G29970	0.470634991	-0.006739502
8	Average	AT3G28960	-0.305421974	-0.012426361
9	A_84_P529752	AT3G28960	0.055673068	0.003263003
10	A_84_P840305	AT3G28960	-0.666517016	-0.028115724
11	Average	AT4G38680	-0.155715316	0.002202385
12	A_84_P829226	AT4G38680	0.036936876	0.004036983
13	A_84_P19571	AT4G38680	-0.578437535	-0.003865307
14	A_84_P804268	AT4G38680	0.07435471	0.006435479
15	Average	AT4G21250	-0.175848445	-3.12E-04
16	A_84_P563561	AT4G21250	-0.456476256	-0.003182524
17	A_84_P790767	AT4G21250	0.104779366	0.002558785
18	Average	AT4G34460	-0.010683424	0.002556537

ログファイルには、一つの遺伝子に複数のプローブが対応していた場合に、出力ファイルに書き出された平均化データ（Average と書かれた行）と、平均計算のもととなったプローブごとのデータが出力されます。
どのプローブのデータを採用すべきかを検討したり、またこれをもとに ID 対応表を詳細に編集したりする際にお役立てください。

2-2. プローブ ID - 遺伝子 ID 対応表の作成支援機能

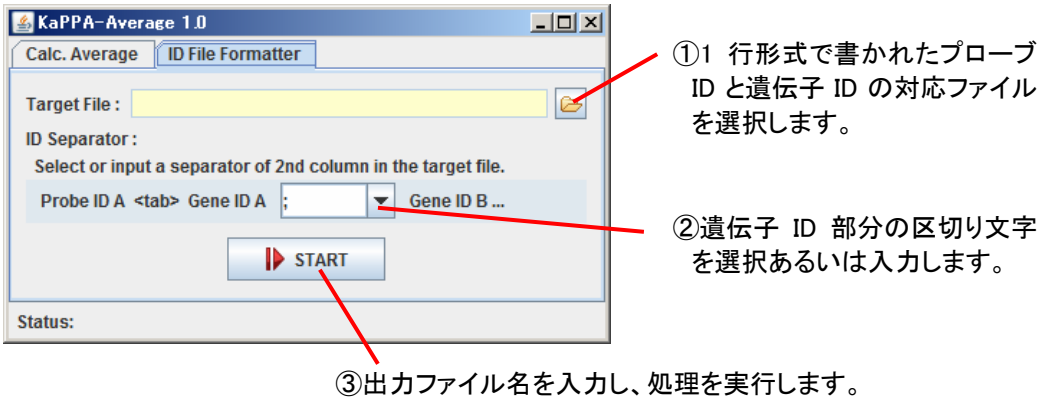
「ID File Formatter」のタブでは、「Calc. Average」で使用するプローブ ID と遺伝子 ID の対応表を作成するための支援機能を提供しています。

マイクロアレイのメーカーなどが提供する情報ファイルには、2-1 で示したようなフォーマットとは別の形で、プローブと遺伝子の対応関係が記されている場合があります。例えば、各プローブ ID に対応する遺伝子 ID が、セミコロンやスペースなどで区切られた文字列として、1 行で書かれている場合があります。

	A	B	C	D	E	F
1	array_element_name	array_element_ty	organism	is_control	locus	description
2	A_84_P10173	oligonucleotide	Arabidopsi	no	AT5G10880	tRNA synthetas
3	A_84_P534845	oligonucleotide	Arabidopsi	no	AT2G01390	pentatricopeptid
4	A_84_P712221	oligonucleotide	Arabidopsi	no	AT1G72860;AT1G72855	[AT1G72860, di
5	A_84_P15539	oligonucleotide	Arabidopsi	no	AT3G24350	SYP32 (SYNTA
6	A_84_P756685	oligonucleotide	Arabidopsi	no	AT2G14370	transposable ele
7	A_84_P820213	oligonucleotide	Arabidopsi	no	AT1G20560	AAE1 (ACYL AC
8	A_84_P18685	oligonucleotide	Arabidopsi	no	AT5G11510	MYB3R-4 (myb
9	A_84_P19533	oligonucleotide	Arabidopsi	no	AT4G29060	emb2726 (embr

例）TAIR が提供する Agilent 社のプローブ ID と AGI コードとの対応関係表

ID File Formatter では、このように 1 行で書かれたデータを、Clac. Average で使用できる複数行形式に変換することができます。



2-2-1. データの準備

もととなるデータは、以下のようなフォーマットとして保存してください。

	A	B
1	A_84_P10173	AT5G10880
2	A_84_P534845	AT2G01390
3	A_84_P712221	AT1G72860;AT1G72855
4	A_84_P126091	AT1G49430
5	A_84_P845289	AT3G10360
6	A_84_P754326	AT1G22065
7	A_84_P552996	AT5G02010
8	A_84_P755542	AT2G15450;AT2G26620;AT2G15470;AT2G15460
9	A_84_P827433	AT2G40120
10	A_84_P762845	AT3G13061;AT3G13060
11	A_84_P12829	AT4G02050
12	A_84_P860413	AT2G43760
13	A_84_P805716	AT5G10430
14	A_84_P206908	AT1G65380

1 列目： プローブ ID
2 列目： 遺伝子 ID が適当な区切り文字で区切られたもの。

ファイルはタブ区切りテキストとして保存してください。

また以下のように、2 列目以降の遺伝子 ID がタブで区切られ 3 列以上が存在するようなたようなファイルも処理することができます。このような 3 列以上が存在するファ

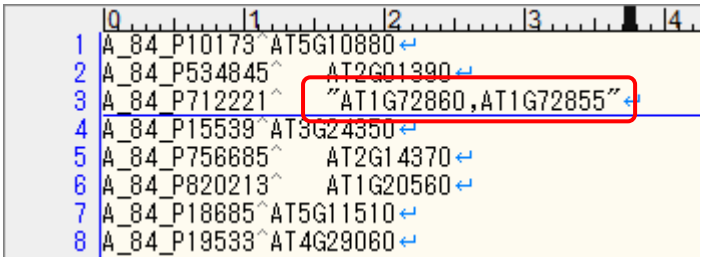
イルは、区切り文字の選択で「<tab>」を指定した時のみ処理可能です（後述）。

	A	B	C	D	E
1	A_84_P10173	AT5G10880			
2	A_84_P534845	AT2G01390			
3	A_84_P712221	AT1G72860	AT1G72855		
4	A_84_P278620	AT3G06960			
5	A_84_P756474	ATMG00513	AT2G07711		
6	A_84_P751911	AT1G66990	AT1G66980		
7	A_84_P312113	AT5G20260			
8	A_84_P23438	AT5G23150			
9	A_84_P755542	AT2G15450	AT2G26620	AT2G15470	AT2G15460
10	A_84_P856377	AT1G72900			
11	A_84_P557966	AT2G33796			
12	A_84_P124342	AT5G59080			
13	A_84_P759125	AT3G30218			

注意

・ヘッダー行はありません。

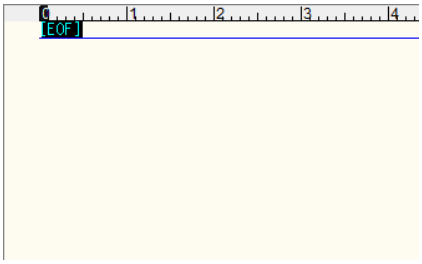
・区切り文字にカンマが使用されている場合、Microsoft の Excel でタブ区切りテキストとして保存すると、下図のように、両脇にダブルクォーテーションが挿入されることがあります。このようなデータは適切に処理されません。



例）Excel で「タブ区切りテキストとして保存」したファイルを、テキストエディタで開いたところ

これを防ぐには、以下のように、テキストエディタで空のテキストファイルを作成し、このなかに Excel のセルをコピーしてペーストします。

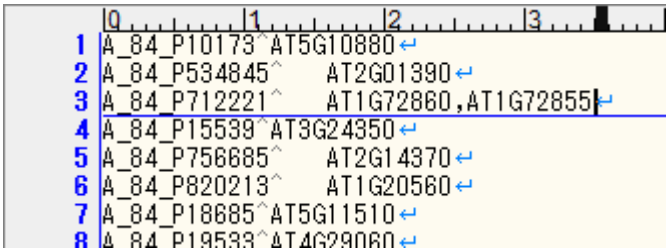
テキストエディタで、新規ファイルを作成。



Excel でデータ部分をコピー

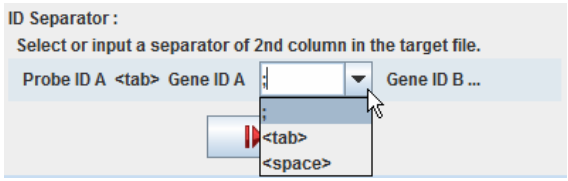
	A	B	C
1	A_84_P10173	AT5G10880	
2	A_84_P534845	AT2G01390	
3	A_84_P712221	AT1G72860,AT1G72855	
4	A_84_P15539	AT3G24350	
5	A_84_P756685	AT2G14370	
6	A_84_P820213	AT1G20560	
7	A_84_P18685	AT5G11510	
8	A_84_P19533	AT4G29060	
9	A_84_P586827	AT5G08185	
10	A_84_P21291	AT3G54300	
11	A_84_P14903	AT5G11920	
12	A_84_P539068	AT4G39290	
13	A_84_P755862	AT2G07480	
14	A_84_P175641	AT3G19660	
15	A_84_P818748	AT3G13560	
16	A_84_P727905	AT1G14650	

テキストエディタに貼り付け



2-2-2. 実行

①「Target File」欄で 2-2-1 で準備したファイルを選択します。
②「ID Separator」の部分で、2 列目の遺伝子 ID の区切り文字を選択、あるいは入力します。

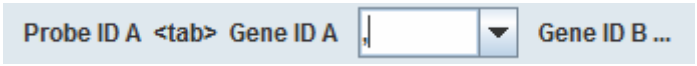


<tab>は、2 列目以降の遺伝子 ID がすべてタブ区切りとして保存された 3 列以上のファイルに対して用います。

<space>は、遺伝子 ID の区切り文字として、一つ以上の半角スペースが用いられている場合に使用します。

その他の区切り文字として、デフォルトではセミコロン（;）が選択できます。他の区

切り文字にしたい場合には、ここに区切り文字を入力してください。



例）カンマ（,）を設定した例

③「START」ボタンを押すと、出力ファイル名を問い合わせるダイアログボックスが現れ、ファイル名を入力すると処理が始まります。Status の欄に「Finished.」と表示されれば、処理は終了です。

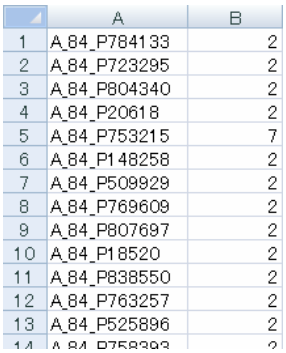
2-2-3. 出力データ

出力データは、2-1 で紹介したような形式となっています。区切り文字で区切られていた遺伝子 ID は、複数行に書き出されています（下図の 3 行目、4 行目を参照）。

	A	B
1	A_84_P10173	AT5G10880
2	A_84_P534845	AT2G01390
3	A_84_P712221	AT1G72860
4	A_84_P712221	AT1G72855
5	A_84_P15539	AT3G24350
6	A_84_P756685	AT2G14370
7	A_84_P820213	AT1G20560
8	A_84_P18685	AT5G11510
9	A_84_P19533	AT4G29060
10	A_84_P586827	AT5G08185
11	A_84_P21291	AT3G54300
12	A_84_P14903	AT5G11920
13	A_84_P539068	AT4G39290
14	A_84_P755862	AT2G07480
15	A_84_P175641	AT3G19660
16	A_84_P818748	AT3G13560
17	A_84_P727905	AT1G14650
18	A_84_P14847	AT4G40060

区切り文字で区切られていた文字列は、すべて遺伝子 ID として書き出されます。もとになるデータによっては、プローブが対応する遺伝子がない場合に「no_match」などと記載されていることがありますが、このような文字列は、除去されずに残っています。出力ファイルを得た後は、適切な対応関係が書き出されているかどうかを必ずチェックし、必要があれば手作業で除去してください。

出力ファイル名の拡張子が「.log」となっているファイルも同時に出力されます。
（例えば、出力ファイル名が「id_formatted.txt」であれば、「id_formatted.log」）



	A	B
1	A_84_P784133	2
2	A_84_P723295	2
3	A_84_P804340	2
4	A_84_P20618	2
5	A_84_P753215	7
6	A_84_P148258	2
7	A_84_P509929	2
8	A_84_P769609	2
9	A_84_P807697	2
10	A_84_P18520	2
11	A_84_P838550	2
12	A_84_P763257	2
13	A_84_P525896	2
14	A_84_P758393	2

このログファイルには、一つのプローブ ID が複数の遺伝子 ID に対応していた場合、その個数が書き出されています。データの解釈において、そのプローブを採用するかどうかの判断材料としてご活用ください。