

Gene Clustering Algorithm GRANK

DNA microarrays allow quantitative measurement of the proportional representation of thousands of genes in a biological sample.

DNA microarrays can be used to measure changes in gene expression

1. Over collections of related samples, such as clinical samples taken from different cancer patients,
2. During a biological process, such as the reaction of a patient to chemotherapy.

Algorithm GRANK

The aim is to identify distinct sets of genes with similar expression patterns across DNA microarrays. The present algorithm basically comprises three major steps. Firstly, the gene with the largest variance is selected. Next, rank correlation is applied to select all genes with expression patterns coherent with that one gene. Finally, with a cluster thus found the search is continued among the genes least correlated.

Accordingly the present algorithm runs as follows:

1. Start with the entire expression matrix, each row centered to have a zero mean.
2. Determine the row of largest variance.
3. Compute for each row the Spearman rank-order correlation coefficient with the one row and select those genes to belong to the cluster that meet a statistical significance level, chosen a priori.
4. Orthogonalize each row of the expression matrix with respect to the average gene in the cluster.
5. Repeat steps 1 and 2 with the orthogonalized data and repeat step 3, to find a second cluster. This process is continued until the vector space spanned by the genes is fully covered.

Use of the Spearman rank-order correlation method ensures that only those genes are selected that are really correlated, that is, to the significance level chosen. Accordingly the clusters will be relatively small: they comprise all genes, *including those of small variance*, that are positively or negatively correlated with the one gene found by shaving. The clusters may in principle overlap. The number of different clusters found is typically less than the number of samples.

The algorithm allows assessment of cluster stability, by varying the significance level. Noise on the gene data could give rise to spurious clustering. If so, consider to rerun the program on data with noise added.

Program considerations

Input

Computer program GRANK starts with stating its name, the name of the author, and a reference to the GRANK End-User License Agreement (see below). It then proceeds with asking three questions:

1. “Enter the file name with the microarray data”
The program expects the data to be offered in a flat file, with the data presented in matrix form. The first column of the matrix must contain the names of the genes (max. 15 characters); the other columns must contain the gene expression values sample by sample. Each row of the matrix therefore contains the name of a gene followed by the expression values of that gene over the biological samples involved. Only blanks or tabs may separate the columns. Each column must have the name of the relevant sample (max. 15 characters) in its first row.
2. “Enter the number of samples (between 3 and 50)”
The program expects an integer number greater than or equal to three. One can set this number lower than the number of samples in the input file, e.g. to distinguish between cancer classes, but one should set the number preferably higher than three for better program performance.
3. “Enter the acceptable clustering error per gene (%)”
This sets the statistical significance level chosen a priori. A clustering error per gene of say 5% means that the gene will be wrongly clustered about once in $20 \cdot M$ trials where M is the total number of genes. It makes sense to start with such a low percentage and then to continue with higher percentages for an impression of more extended gene networks.

The current implementation of program GRANK allows for a maximum of 4050 genes and a maximum of 50 biological samples.

It is imperative that the input file matrix has no missing entries; zero is a valid entry. Failure to meet this requirement will lead to an error message and stop further execution of the program.

Output

The program output is presented in a flat file ARROUT.TXT that it creates when absent. The program first recites some input data: the title of the program, the name of the input file, the numbers of biological samples and genes involved, and the chosen clustering error per gene.

Next the program writes a table for each different cluster identified. The table comprises: the names of the samples in the first row, the names of the genes in the first column and their expression values centred to have zero mean in subsequent columns, and a +/- sign in the last column. This sign denotes whether the gene correlates (+) or anti-correlates (-) with the other genes in the cluster and whether the expression values have their sign flipped accordingly. The table is preceded by the cluster number and the total number genes in the cluster, and is followed by the cumulative gene vector space covered.

The tables can be directly imported as text into, for instance, an MS Worksheet for further processing such as graphing or colour coding of the gene expression values.

The program keeps track of progress with screen messages "... cluster xx identified ..." and ends with screen message "Program has finished" when finished.

July 2002

Example input

	Sample1	Sample2	Sample3	Sample4	Sample5
Ntfn2__	-2.11E-01	8.43E-02	-9.48E-02	-3.10E-01	-6.02E-03
Berp__	7.83E-02	1.72E-01	-5.27E-02	2.41E-02	6.02E-03
CARP__	-4.49E-01	-1.69E-01	-1.10E-01	-4.21E-02	1.81E-01
cfos__	-1.17E-01	2.47E-01	5.27E-02	-6.32E-02	-1.05E-01
NAC-1__	-2.38E-01	5.72E-02	8.58E-02	8.81E-02	-1.02E-01
bgCREB__	-8.13E-02	7.98E-02	-2.86E-02	-6.62E-02	7.53E-03
cjun__	5.72E-02	2.86E-02	-9.78E-02	2.41E-02	-7.53E-03
Krox-20__	-1.29E-01	4.36E-02	2.86E-02	-3.09E-01	-2.62E-01
Krox-24__	-2.65E-01	-4.14E-01	1.94E-01	-4.52E-03	-1.23E-01
GlucortR__	-1.84E-01	3.06E-01	3.15E-01	4.36E-02	0.00E+00
Hes-1__	-3.10E-01	5.27E-02	8.58E-02	-6.02E-02	2.26E-02
junD__	1.99E-01	4.36E-02	4.36E-02	1.23E-01	1.96E-02
etc.

Example output

PROGRAM GRANK 4050 x 50

Input file processed: Arrin.txt
Number of samples: 5
Number of genes: 318
Clustering error per gene: 5.00 %

Cluster_01: 10 genes

	Sample1	Sample2	Sample3	Sample4	Sample5	
cfos__	1.20E-01	-2.44E-01	-5.00E-02	6.59E-02	1.08E-01	-
AChRa6__	2.33E-01	-1.92E-01	-2.00E-02	-1.10E-02	-1.03E-02	+
NR2b__	1.66E-01	-2.24E-01	-4.03E-02	2.59E-02	7.25E-02	+
EphRA4__	3.25E-02	-3.52E-02	-6.62E-03	-6.02E-04	9.93E-03	+
Nogo-b__	7.10E-02	-7.95E-02	-4.18E-02	-1.78E-02	6.80E-02	-
NSE__	1.27E+00	-3.65E-01	-3.47E-01	-2.83E-01	-2.78E-01	+
Homer-1a__	3.09E-01	-4.54E-01	-1.38E-01	1.15E-01	1.68E-01	-
Chat__	2.13E-01	-2.49E-01	-5.36E-02	-1.14E-02	1.01E-01	-
SHPS-1__	7.07E-02	-6.47E-02	-3.76E-02	-1.51E-02	4.67E-02	+
PKAreg1__	7.53E-02	-6.77E-02	-3.76E-02	-7.53E-03	3.76E-02	-

Vector space covered: 32 %

Cluster_02: 6 genes

	Sample1	Sample2	Sample3	Sample4	Sample5	
Wnt7A__	-7.41E-02	1.93E-02	-3.64E-02	1.02E-01	-1.08E-02	+
CAMK2a__	-9.87E-02	5.33E-02	-3.25E-02	1.03E-01	-2.50E-02	+
PKAcat__	-1.05E-01	4.67E-02	-1.51E-02	6.02E-02	1.35E-02	+
PKC1/2__	-7.97E-02	3.12E-02	-4.04E-04	4.48E-02	4.11E-03	+
b-actin__	-7.47E-01	5.30E-01	-6.90E-01	5.73E-01	3.34E-01	+
GluR5__	-1.51E-01	1.48E-02	-1.69E-02	1.53E-01	-3.01E-04	+

Vector space covered: 51 %

Cluster_03: 7 genes

	Sample1	Sample2	Sample3	Sample4	Sample5	
Rac1_____	6.02E-04	6.62E-03	5.12E-03	7.13E-02	-8.37E-02	+
GABARb2	-8.91E-02	7.50E-02	6.44E-02	1.56E-01	-2.07E-01	+
IP3-R_____	-8.28E-02	1.14E-01	-4.21E-02	1.44E-01	-1.34E-01	+
OBCAM_____	-1.75E-02	7.74E-02	5.03E-02	1.23E-01	-2.33E-01	+
bFGF_____	6.02E-04	2.62E-02	5.12E-03	1.06E-01	-1.38E-01	-
GAP43_____	-1.57E-02	1.24E-01	4.91E-02	3.71E-01	-5.29E-01	+
Narp_____	1.81E-02	2.86E-02	2.71E-02	2.15E-01	-2.89E-01	+

Vector space covered: 76 %

Cluster_04: 9 genes

	Sample1	Sample2	Sample3	Sample4	Sample5	
MAOb_____	-3.97E-02	-6.08E-02	-2.11E-03	7.62E-02	2.65E-02	+
GABARg3__	-2.38E-02	-1.23E-01	4.18E-02	5.52E-02	4.92E-02	+
egr-3_____	6.92E-02	-4.95E-01	9.63E-02	2.00E-01	1.29E-01	+
5HT1dR_____	-9.90E-02	-1.50E-01	1.81E-03	1.94E-01	5.30E-02	-
NR1_____	-6.95E-02	-1.21E-01	3.13E-02	8.10E-02	7.80E-02	+
GABARd_____	7.75E-03	-4.79E-02	1.08E-02	1.57E-02	1.38E-02	+
GABARe_____	-2.71E-02	-1.37E-01	-9.03E-03	1.22E-01	5.12E-02	+
b-actin_____	-2.58E-01	-3.20E-01	-2.25E-01	9.67E-01	-1.64E-01	+
GAPDH_____	-3.58E-02	-1.63E-01	-1.80E-02	1.96E-01	2.02E-02	+

Vector space covered: 100 %

GRANK End-User License Agreement

Computer program GRANK is made available to the scientific community free of charge in the interest of scientific progress. Ownership of the program and associated documentation, however, remains with the author to which all queries should be addressed:

Dr. W.P.H. de Boer
Vrije Universiteit medical centre, Dept. Medical Oncology, BR232
De Boelelaan 1117
NL-1081 HV AMSTERDAM, The Netherlands
Email: WPH.deBoer@VUmc.nl

GRANK has been designed to identify distinct sets of genes with similar expression patterns across DNA micro-arrays. Genes are clustered using rank correlation to ensure that the genes selected are really correlated, that is, to a significance level chosen. Clusters will be fairly small: they comprise all genes, including those of small variance, which are positively or negatively correlated. The search for clusters continues until there are no least-correlated genes left with the clusters already found.

IMPORTANT – READ CAREFULLY: By copying, installing or otherwise using computer Program GRANK and associated Documentation, User agrees to be bound by the terms and conditions of this GRANK End-User License Agreement (“Agreement”).

1. **DEFINED TERMS.** For purposes of this Agreement: (a) “Licensor” means W.P.H. de Boer, voluntary research associate at the VU medical centre (“VUmc”) in Amsterdam, the Netherlands; (b) “Program” means computer program GRANK; (c) “Documentation” means the user manual associated with the Program; (d) “User” means the user of the Program and associated Documentation; and (e) “VU” means the Vrije Universiteit in Amsterdam, the Netherlands.
2. **GRANT OF LICENSE.** Licensor hereby grants User free of charge a non-exclusive, transferable, perpetual license to use the Program in machine-readable form together with associated Documentation subject to the terms and conditions set forth in this Agreement.
3. **RESTRICTIONS.** User may not de-compile or disassemble the Program. User may not rent, lease or otherwise merchandize the Program.
4. **COPYRIGHT.** User understands and agrees that Licensor is not selling or transferring in whole or in part any title to Program and associated Documentation. User may permanently transfer all User’s rights under this Agreement, provided the recipient agrees to the terms of this Agreement.
5. **NO WARRANTY.** Licensor expressly disclaims any warranty for the Program. The Program and any related Documentation is provided “as is” without warranty of any kind, either expressed or implied. The entire risk arising out of use or performance of the Program remains with the User.
6. **LIMITATION OF LIABILITY.** In no event will Licensor, VU or VUmc be liable to User for damages, including any general, special, incidental, or consequential damages arising out of the use of the Program.
7. This Agreement shall be exclusively interpreted in accordance with and governed by the Netherlands’ Dutch law, in particular the Copyright Act (Auteurswet).

July 2002