Progression Analysis of Disease *PAD*

web tool for the paper *:

Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival

M.Nicolau, A.Levine, G.Carlsson

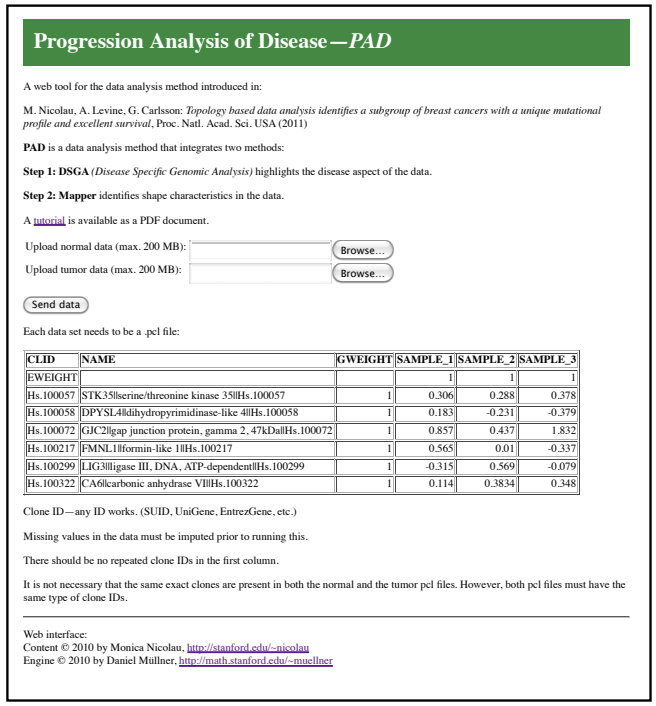
Proc Natl Acad Sci U S A (2011)

TUTORIAL

for using the web interface

code M.Nicolau & G.Singh
web engine D.Müllner
tutorial M. Nicolau

* ***PAD*** is free to all who want to try it. We ask only that you please reference the paper if you use ***PAD***.



Progression Analysis of Disease—*PAD*

A web tool for the data analysis method introduced in:

M. Nicolau, A. Levine, G. Carlsson: *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, Proc. Natl. Acad. Sci. USA (2011)

PAD is a data analysis method that integrates two methods:

Step 1: DSGA *(Disease Specific Genomic Analysis)* highlights the disease aspect of the data.

Step 2: Mapper identifies shape characteristics in the data.

A tutorial is available as a PDF document.

Upload normal data (max. 200 MB): Browse...

Upload tumor data (max. 200 MB): Browse...

Send data

Each data set needs to be a .pcl file:

CLID	NAME	GWEIGHT	SAMPLE_1	SAMPLE_2	SAMPLE_3
EWEIGHT			1	1	1
Hs.100057	STK35|serine/threonine kinase 35|Hs.100057	1	0.306	0.288	0.378
Hs.100058	DPYSL4|dihydropyrimidinase-like 4|Hs.100058	1	0.183	-0.231	-0.379
Hs.100072	GJC2|gap junction protein, gamma 2, 47kDa|Hs.100072	1	0.857	0.437	1.832
Hs.100217	FMNL1|formin-like 1|Hs.100217	1	0.565	0.01	-0.337
Hs.100299	LIG3|ligase III, DNA, ATP-dependent|Hs.100299	1	-0.315	0.569	-0.079
Hs.100322	CA6|carbonic anhydrase VI|Hs.100322	1	0.114	0.3834	0.348

Clone ID—any ID works. (SUID, UniGene, EntrezGene, etc.)

Missing values in the data must be imputed prior to running this.

There should be no repeated clone IDs in the first column.

It is not necessary that the same exact clones are present in both the normal and the tumor pcl files. However, both pcl files must have the same type of clone IDs.

Web interface:
Content © 2010 by Monica Nicolau, http://stanford.edu/~nicolau
Engine © 2010 by Daniel Müllner, http://math.stanford.edu/~muellner

Progression Analysis of Disease (PAD) – Web Tool Tutorial

Page 1: Upload data files

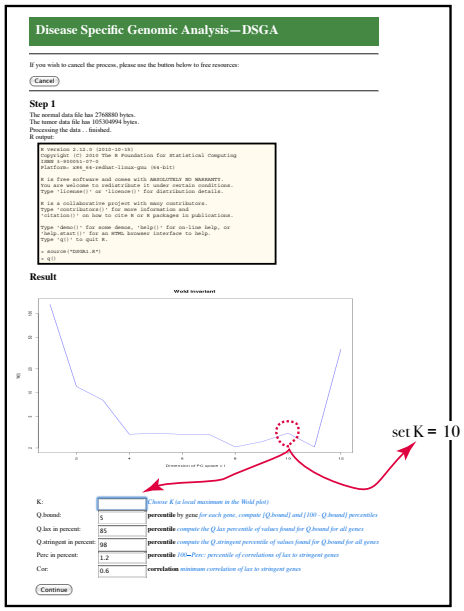
You must upload 2 pcl files, one for the diseased tissue data, a second for the normal tissue data.

- These are not interchangeable. Although the mathematics will work if you switch disease and normal tissue files, the model will be biological nonsense.
- Data must have no missing values. Use, for example a *knn*-impute algorithm.
- Data must be in the form of a standard *pcl* file, as shown to the left.

The clone identifiers in the first CLID column can be any accepted type, but:

they must be the same type for normal and for disease data

there must be no repeats in this CLID column - each ID must occur only once.



Progression Analysis of Disease (PAD) - Web Tool Tutorial

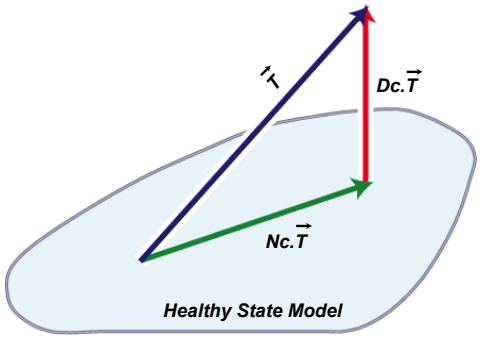
PAD Part 1: Perform **Disease Specific Genomic Analysis** (DSGA)

Page 2: DSGA

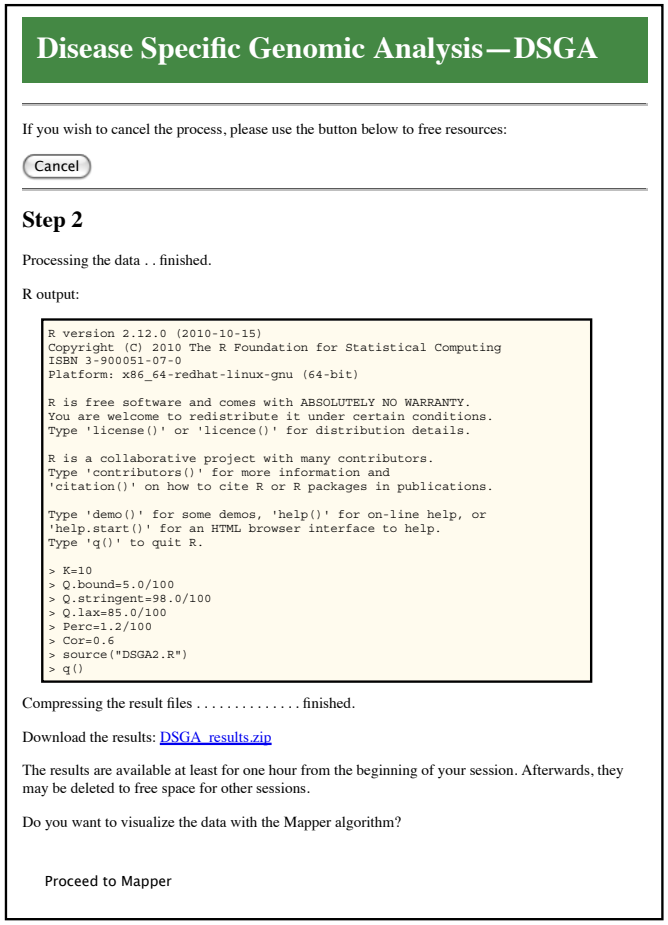
Three steps are performed on this page:

Step 1: construct the mathematical **Healthy State Model (HSM)**
This builds a space from normal tissue data, but you there is a dimension reduction part: you much choose a dimension which gives good signal - to - noise.
In the Wold graph shown, choose a value K for which the plot jumps up (*see screenshot*)

Step 2: Diseased tissue data is transformed to measure deviation from the HSM. This happens in the background but it is the central transformation of DSGA.



Step 3: Gene threshholding on the DSGA-transformed data set.
Roughly genes are retained if they deviate from normal significantly.
You choose parameters for the thresholding roughly as follows:
Q.bound: For each gene take the larger in absolute value of
100 - Q.bound percentile and *Q.bound* percentile Denote by **Q**
Q is a measure of how much a given gene deviates from Healthy in either positive or negative direction
Compute the values for **Q** for all genes then threshhold genes by 2 threshholds:
stringent genes: their **Q**-value is in the highest *Q.stringent* percentile
lax genes: their **Q**-value is in the highest *Q.lax* percentile
Retain lax genes that are highly correlated ($R > Cor$) to stringent genes.
It is reasonable to use default values



Disease Specific Genomic Analysis—DSGA

If you wish to cancel the process, please use the button below to free resources:

Cancel

Step 2

Processing the data . . finished.

R output:

```
R version 2.12.0 (2010-10-15)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> K=10
> Q.bound=5.0/100
> Q.stringent=98.0/100
> Q.lax=85.0/100
> Perc=1.2/100
> Cor=0.6
> source("DSGA2.R")
> q()
```

Compressing the result files . . . . . finished.

Download the results: DSGA_results.zip

The results are available at least for one hour from the beginning of your session. Afterwards, they may be deleted to free space for other sessions.

Do you want to visualize the data with the Mapper algorithm?

Proceed to Mapper

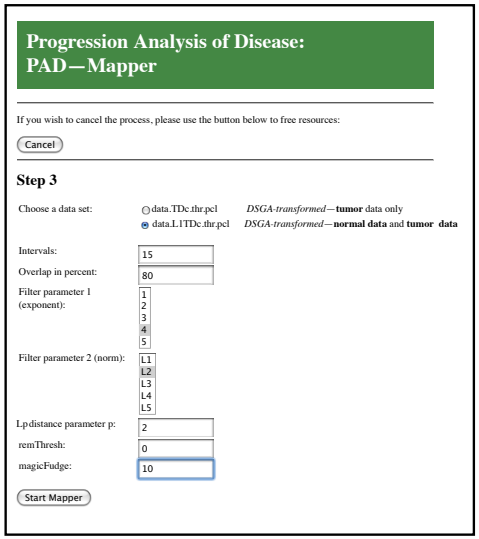
Progression Analysis of Disease (PAD) - Web Tool Tutorial

PAD Part 1: Perform **Disease Specific Genomic Analysis** (DSGA)

Page 3: DSGA

Data transformation and gene threshholding
are performed in the background.
At the end, simply download the DSGA analysis output.
It consists of the following files:

data.Tdis.pcl - tumor tissue disease component data
normal.L1out.pcl - estimate of normal data deviation from the model HSM using a lieave-one-out process
normal.Ndis.pcl -normal tissue disease component
normal.NormalModel.pcl - healthy state model data
data.Tnorm.pcl - fit of tumor tissue onto normal tissue model HSM
normal.Nnorm.pcl - fit of normal tissue onto normal tissue model HSM
thrsdhholded pcl files with reduced number of genes:
data.TDc.thr.pcl : disease component of tumors (as data.Tdis.pcl) with fewer genes: only retained genes that passed threshholds in Page 2
data.L1TDc.thr.pcl : disease component of tumors and normal estimate
data.TDc.AGmc.thr.pcl : disease component of tumors, Array and Gene mean-centered.
data.L1TDc.AGmc.thr.pcl : disease component of tumors and normal estimate, Array and Gene centered at the mean of the tumors.
wold.png is the wold plot used to choose K
parameters.txt record of your choices
record.doc full record of DSGA analysis.



Progression Analysis of Disease (PAD) - Web Tool Tutorial

PAD Part 2: Run *Mapper* on DSGA-transformed data with filter functions defined by DSGA.

Page 4: Mapper:

Choose data and mapper parameters:

Data:

data.TDc.thr.pcl = only tumor tissue data

data.L1TDc.thr.pcl = tumor and normal tissue data

Mapper parameters (try several)

Intervals: number of intervals to subdivide the mapper plot

Overlap: percentage of overlap between intervals.

The filter on each tumor sample (column of data matrix) measures the size (magnitude) of the column vector. However, magnitude can be measured in several ways, and different Lp magnitudes (L_2 magnitude is the standard euclidean distance) together with several powers of these magnitudes, in essence provide different smooth stretches of the graph that is the output of Mapper. Mapper also takes into account how close or far the data points are from one another, and for this Mapper can use different distances, for which the user can choose "Lp distance parameter". The local clustering relies on a histogram of the data, and MagicFudge provides a measure of number of breaks or subintervals in the histogram.

Finally, the Mapper output can be more easily read and interpreted if bins with few points are omitted (remove bins with fewer than *remThresh* points).

Progression Analysis of Disease: PAD—Mapper

If you wish to cancel the process, please use the button below to free resources:

Cancel

Step 4

Processing the data . finished.

R output:

```
R version 2.12.0 (2010-10-15)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> source("Filter.R")
> q()
```

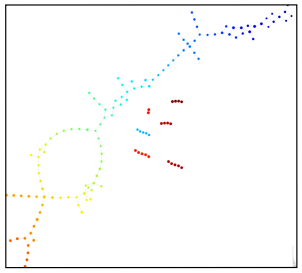
Processing the data . . finished.

MATLAB output:

```
Warning: No window system found.  Java option 'MWT' ignored

                         < M A T L A B (R) >
               Copyright 1984-2008 The MathWorks, Inc.
                  Version 7.7.0.471 (R2008b)
                            September 17, 2008


  To get started, type one of these: helpwin, helpdesk, or demo.
  For product information, visit www.mathworks.com.

>> >> Mapper parameters:
dataset: 'data.L1TDc.thr.pc1'
intv1: 15
ovlap: 80.000000
filter1_set_num: 4
filter1_column: 2
distp: 2
resThresh: 0
magicFudge: 10

Mapper : Filter Range [3735.27-15118051.64]
Mapper : Interval Length : 1007621.09
```

Run graphviz (pdf) . (png) . . . . . (resize) . . (clean up) finished. Compressing the result files . . . finished.

Result



View the graph as a PDF file: PDF

Download the results: Mapper_results.zip.

Here are the results from the DSGA (Steps 1 and 2) again. You might have downloaded them already afte
DSGA_results.zip
The results are available for at least one hour from the beginning of your session. Afterwards, they may be deleted to free

space for other sessions.

Do you want to run the Mapper algorithm again, with different parameters?

Return to Mapper

Please click the button below when you have finished analyzing and downloading the data. This frees resources for your next data set or for other users.

Finish

Progression Analysis of Disease (PAD) - Web Tool Tutorial

PAD Part 2: Run *Mapper* on DSGA-transformed data with filter functions defined by DSGA.

Page 5: Mapper:

Mapper runs of data along chosen parameters.
Output graph is seen on the screen.
Output can be downloaded.

Note that bins are numbered, and the file:
....seqprog_output.txt gives a list of all the sample points sitting in each bin.

After downloading the output, it is a good idea to run the Mapper part several times, with different parameters.