GenePattern

NMFConsensus Documentation

Description: Non-negative Matrix Factorization Consensus Clustering
Author: Pablo Tamayo (Broad Institute) gp-help@broad.mit.edu with contributions from Jean-Philippe Brunet and Ted Liefeld.

Summary: Non-negative matrix factorization (NMF) is an unsupervised learning algorithm [1] that has been shown to identify molecular patterns when applied to gene expression data [2]. Rather than separating gene clusters based on distance computation, NMF detects context-dependent patterns of gene expression in complex biological systems.

The basic principle of dimensionality reduction via matrix factorization operates as follows: given an N x M data matrix A with non-negative entries, the NMF algorithm iteratively computes an approximation, A ~ WH, where W is an N x k matrix, H is a k x M matrix, and both are constrained to have positive entries. For DNA microarrays, N, the number of genes, is typically in the thousands. M, the number of experiments, rarely exceeds a hundred, while k, the number of classes to be determined depends on the heterogeneity of the dataset. The algorithm starts with randomly initialized matrices of the appropriate size, W and H. These are iteratively updated to minimize the Euclidean distance between V and WH or a divergence norm [3]. The program also computes row and column factor memberships according to maximum amplitudes. This membership information is also used to sort the output matrices according the row and column membership (the row and columns are then relabeled: <name>_f<NMF factor>.

This version is an R version of the Euclidean and Divergence NMF equations. It is slow and is intended for exploratory use. A faster version of the NMF Consensus is available in MATLAB from the Broad Institute web site. Running NMFConsensus using the settings from Brunet et al (2004)

ALL AML dataset (~5000 genes)
k.initial = 2
k.final = 5
num.clusterings = 20
num.iterations = 2000
stop.convergence=40
stop.frequency=10

takes approximately 2.2 hours to complete on a Pentium M 1.8GHz processor.

References:

1. Lee, D.D and Seung, H.S. (1999), ‘Learning the parts of objects by non-negative matrix factorization’, Nature 401, 788-793.
2. Jean-Philippe Brunet, Pablo Tamayo, Todd Golub, Jill Mesirov (2004). Matrix Factorization for Molecular Pattern Recognition, PNAS 101, 4164-4169.
3. Lee, D.D., and Seung, H.S., (2001), ‘Algorithms for Non-negative Matrix Factorization’, Adv. Neural Info. Proc. Syst. 13, 556-562.

Parameters:

Name	**Description**
dataset.filename	Input dataset (gct or res)
k.initial	Initial value of K**.** (Default: 2)
k.final	Final value of K. (Default: 5)
num.clusterings	Number of clusterings to perform for each value of K**.** (Default: 20)
max.num.iterations	The maximum number of iterations to perform for each clustering run for each value of K. This number may not be reached depending on the stability of the clustering solution and the settings of stop convergence and stop frequency. (Default: 2000)
error.function	The error function to use (divergence or euclidean) (Default: divergence)
random.seed	Random seed used to initialize W and H matrices by the random number generator. e.g. 4585, 4567, 5980. This may be set to provide repeatable results for given parameter inputs even though the algorithm is properly random. (Default: 123456789)
output.file.prefix	Prefix to prepend to all output file names.
stop.convergence	How many "no change" checks are needed to stop NMF iterations before max iterations is reached (convergence). Iterations will stop after this many "no change" checks report no changes. (Default: 40)
stop.frequency	Frequency of "no change" checks. NMFConsensus will check for changes every 'stop frequency' iterations. (Default: 10)

Output Files:

1. membership.gct: membership results for samples at all values of K
2. cophenetic.txt: cophenetic values for each K
3. cophenetic.plot.pdf: plot of cophenetic for each value of K
4. consensus.k.#.gct (for each value of K): consensus matrix for k=#
5. consensus.plot.k#.pdf (for each value of K): plot of consensus matrix for k=#
6. graphs.k#.pdf (for each value of K): Plots of the ordered consensus matrix and ordered linkage tree and sample plots of NMF convergence, W matrix, H matrix (ordered and unordered), metagenes (ordered and unordered)

Platform dependencies:

Module type:	Projection
CPU type:	any
OS:	any
Language:	R 2.5

GenePattern

GenePattern Version Notes:

Date	Version	Description
02/19/08	4	Updated for R 2.5
08/17/12	5	Added output files to manifest. Doc edits.