KSscore Documentation

Module name:	KSscore
Description:	Computes the KS score for a set of genes within an ordered list.
Author:	Justin Lamb (Broad Institute) justin@broad.mit.edu
Date:	9/2/03
Release:	1.0

Summary: This task returns a numeric value ('KS score') representing the positional distribution of a set of query genes (*i.e.* 'tags') within an ordered list of genes.
The user supplies the tags in a text file that contains one unique feature identifier per line. The ordered list is supplied in a .pol file. A .pol file is tab-delimited format in which the first column is a ranking, the second column is the unique feature identifier, the third column is a text field (usually a description of the feature), and the last column is a value upon which the rank position is based.
KS score is computed in accordance with the Kolmogorov-Smirnov non-parametric rank statistic where *X* is the number of genes in the query gene set, *Z* is the number of genes in the ordered list, and $Y = Z - X$. A vector *V* is then constructed where *V(i)* is the component corresponding to gene *i* from the ordered list, with *V(i)* = *Y* if gene *i* is in the query gene set and *V(i)* = *–X* if not. Thus,

$$\sum_{i=1}^{Z} V(i) = 0$$

The KS score is defined as the maximum value of the running sum of consecutive values of *V* defined as

$$\max_{j} \sum_{i=1}^{j} V(i)$$

The STDOUT from this task reports the number of genes in the ordered list, the number of genes in the query gene set, the value of *Y* (*i.e.* 'positive score'), and the value of *–X* (*i.e.* 'negative score'). Also returned in a table showing the position of each of the query genes in the ordered list, the value of *V* as that gene, that gene's unique identifier, and its description. The value of *V* at the last gene in the ordered list is reported (*i.e.* 'running sum at the end'), together with the KS score itself. The value provided in parentheses is the minimum value of the running sum of consecutive values of *V*. Note that if the 'running sum at the end' is not zero the KS score reported is not valid (see above). The most likely reason for this is that one or more of the query genes are not included in the ordered list. The output file from this task is a .pol file in which the first column is the rank of each query gene, the second is its unique identifier, the third is the gene description and the fourth is the position of that gene in the ordered list.
The *KSscore* task is used to examine the enrichment of a set of genes at the top of an ordered list. The KS score is high when the tags appear early (*i.e.* near the top) of the ordered list. The significance of the KS score for a particular test may be examined by computing KS scores for multiple sets of *X* query genes selected at random from the dataset (note that the KS score is not independent of the number of members of the query gene set). Alternatively, the query gene set may be kept constant and the ordered list shuffled. Another permutation option—also implemented as a data-mining approach—is to test the enrichment of a query gene set within the ordered lists populated by the nearest neighbors of all genes in a database in turn. This process is referred to as Kolmogorov-Smirnov Scanning (KSS).

References:

- Hollander and Wolfe (1999) Nonparametric Statistical Methods, second edition (Wiley)
- Lamb *et al.* (2003) A Mechanism of Cyclin D1 Action Encoded in the Patterns of Gene Expression in Human Cancer Cell 114: 323-334

Usage/Example:

```
KSscore <-
function(query.filename, output, input.filename, server=defaultServer)
{
return (runAnalysis("KSscore", "query.filename"=query.filename.grp, "output"=output,
"input.filename"=input.filename.pol, server=server))
}
```

Parameters:

Name	**Description**
input filename	the .pol file containing the ordered list
query filename	query gene set file containing one identifier per line
output	name of the output file (a .pol extension will be appended)

Return Value: An R list with components:

1. output.pol: a parametized ordered list reporting the position of the query genes in the ordered list
2. STDOUT: a report containing the number of genes in the ordered list, the number of genes in the query gene set, the value of *Y* (*i.e.* 'positive score'), the value of *–X* (*i.e.* 'negative score'), a table showing the position of each of the query genes in the ordered list, the value of *V* as that gene, that gene's unique identifier, and its description, then the value of *V* at the last gene in the ordered list is reported (*i.e.* 'running sum at the end'), the KS score itself, and the minimum value of the running sum of consecutive values of *V* (in parentheses).
3. STDERR: the standard error report from the program

Platform dependencies:

Task type:	Statistical Methods
CPU type:	any
OS:	any
Java JVM level:	n/a
Language:	Perl
Support files:	none

Native command line: <perl> KSscore.pl <input.filename.pol> <query.filename.tag/.grp> <input.filename_path><file.separator><output>