MedlineR: an Open Source Library in R for Medline Literature Data Mining

Draft 1-21-04

Revised 3-10-04, 5-10-04

Abstract

Summary: We describe an open source library written in the R programming language for Medline literature data mining. This MedlineR library includes programs to query Medline through the NCBI PubMed database; to construct the co-occurrence matrix; and to visualize the network topology of query terms. The open source nature of this library allows users to extend it freely in the statistical programming language of R. To demonstrate its utility, we have built an application to analyze term-association by using only ten lines of code. We provide MedlineR as a library foundation for bioinformaticians and statisticians to build more sophisticated literature data mining applications.

Keywords: literature, text data mining, Medline, PubMed

Availability: The library is available from
http://dbsr.duke.edu/pub/MedlineR

Contact: Lin00025@mc.duke.edu, MCCon012@mc.duke.edu, johns001@mc.duke.edu, shoem003@mc.duke.edu

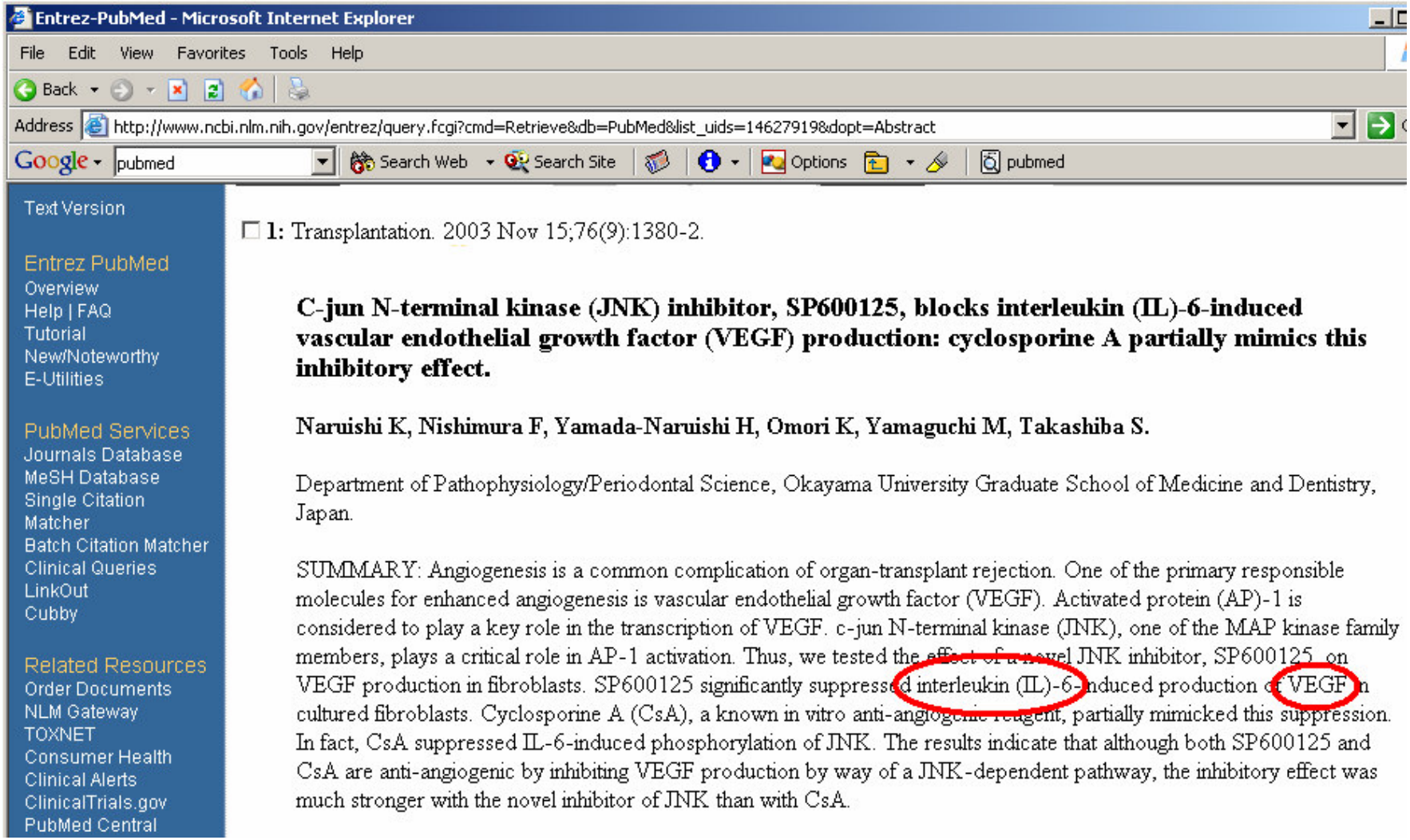
Motivation

- What to do with a list of genes?
  - To move on from a gene list to a biological network: InPharmix product
  - To built a foundation for advanced statistical analysis of genomics literature

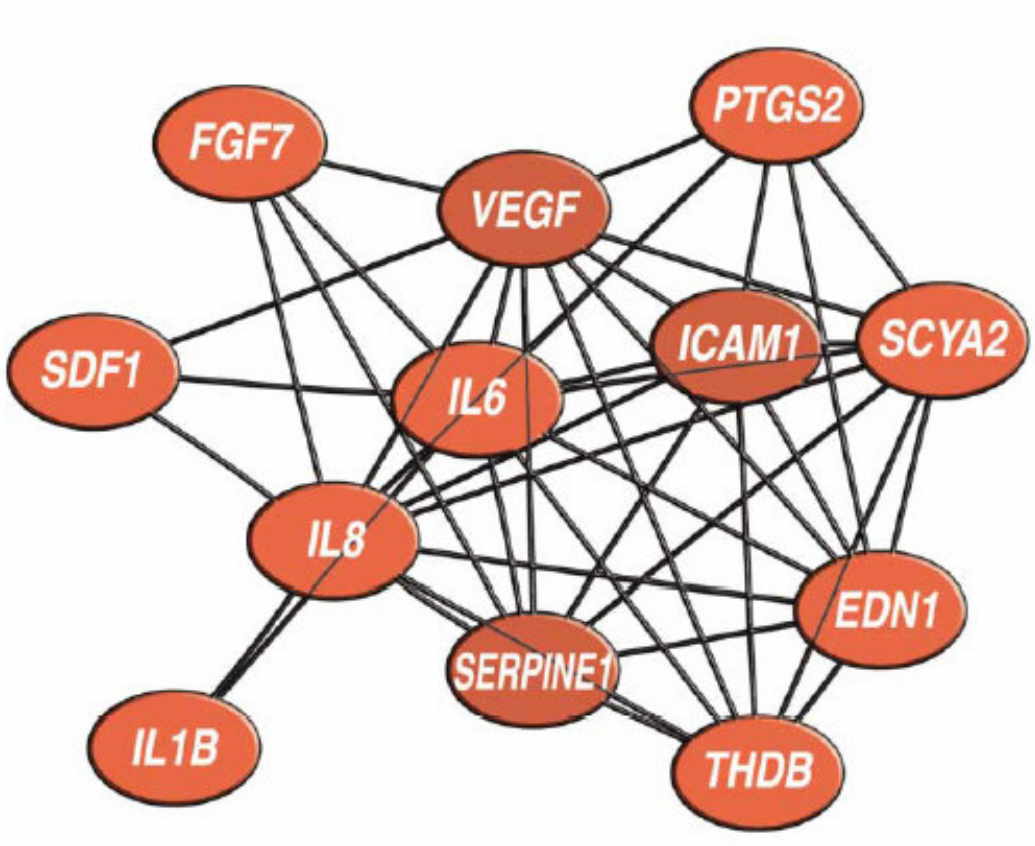
A Comparison of Medline data mining programs

Program	open source	programming control and user extensibility	automatic pairwise query	visualization	reference
PubMed	-	+	-	-	Wheeler et al., 2002
PubMatrix	-	-	+	+	Becker et al., 2003
PubGene	-	-	+	+	Jenssen et al., 2001
Inpharmix	-	-	+	+	Inpharmix Inc., Greenwood, IN
MedlineR	+	+	+	+	this paper

Establish a Connection



Co-occurrence based literature network



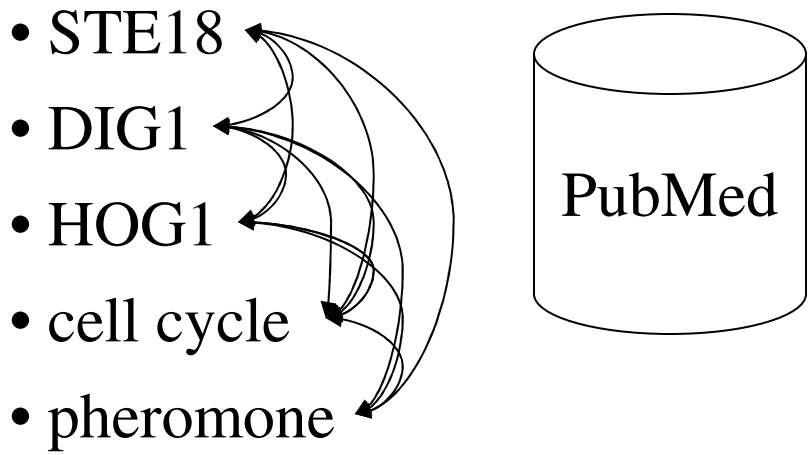
- PubGene. *Nature Genetics* 28:21, 2001

- InPharmix. *CAMDA* II:196, 2001

- PubMatrix. *BMC Bioinformatics* 2003, 4:61

How?

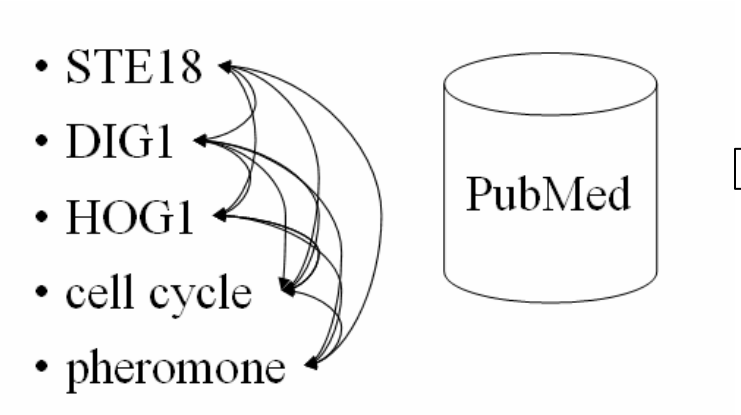
- Starting from a list of genes

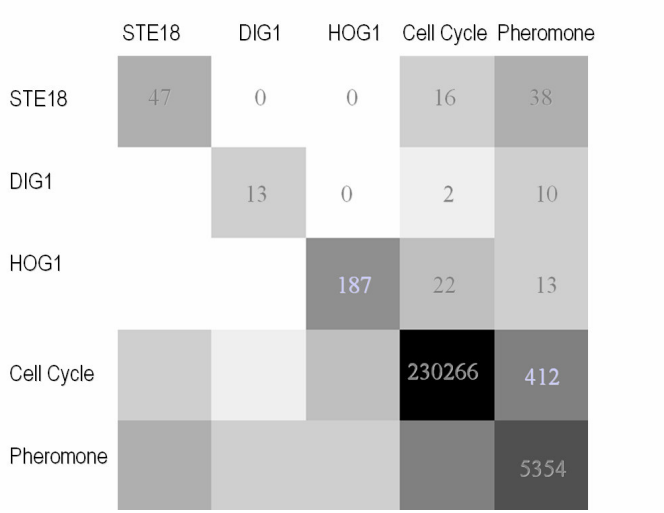


- Want to know their association in the literature

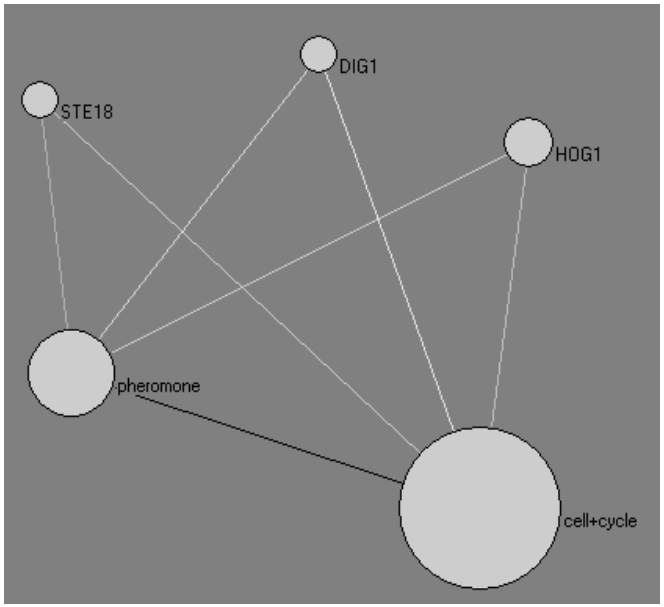
(A) Term list

(B) Association matrix





(C) Network visualization



What is in the library?

Program	Description	Example Usuage
findAssociation	Mining the literature association among a list of terms	findAssociation(termList=c("STE12", "DIG1", "HOG1", "SST2"))
countApair	Count the abstracts with both terms in the pair	countApair(term1="STE18", term2="STE12")
getAmatrix	return a co-occurance matrix for a list of terms	getAmatrix(termList=c("STE12", "DIG1", "HOG1"))
write.pajek	output in Pajek format for visualization	write.pajek(result.matrix, fileName="genes.net")
sleeping.delay	pause betw een queries, accodrding to NCBI	sleeping.delay()
esearch	query NCBI Pubmed	esearch ("term=cancer+OR+diabetes")
fetchAnAbstract	Retrieve a MEDLINE abstract	fetchAnAbstract (pmID=134567)

Open Source Development

Feature	SourceForge	Bioinformatics.org
Homesite	x	x
Discussion forums	x	x
Bug tracking	x	x
Mailing lists	x	x
Surveys		x
CVS	x	x
HTTP download	x	x
Shell account		x
News items	x	x
RSS Feeds	x	
Support requests	x	
Patch tracking	x	
Feature requests	x	
Document manager	x	
Task manager	x	x
Project statistics	x	