Correlation Analysis Cell:
A Tutorial

Version 1.0

Table of Contents

About this Guide

Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH Roadmap National Centers for Biomedical Computing (http://www.bisti.nih.gov/ncbc/). One of the goals of i2b2 is to provide clinical investigators broadly with the software tools necessary to collect and manage project-related clinical research data in the genomics age as a cohesive entity—a software suite to construct and manage the modern clinical research chart.

The guide provides installation steps for the *Correlation Analysis Cell* of the i2b2 hive. This specialized analysis cell uses mutual information theory to calculate observed correlations within the data of the hive. This type of cell represents an important achievement of the hive.

Document Version History

Date	Revision	Description	Author(s)
June 27, 2008	version 1.0	Initial revision, 1.0	Vlad Valtchinov

1

Prerequisites and Third-Party Software

Downloads and Installation

a. i2b2 Workbench version 1.2.3 or 1.3

Download i2b2 Workbench version 1.2.1 (i2b2Workbench-src-121.zip) from https://www.i2b2.org/software/repository.html?t=demo&p=14. Follow installation and configuration instructions as given in the *i2b2 Workbench Developers' Guide v1.2.1* which can be found under the Docs tab.

b. Java JDK 5.0 – needed for i2b2 Workbench

This version of the DJK is needed for running the Eclipse Workbench. Download JDK 5.0 Update 11 (jdk-1_5_0_11-windows-i586-p.exe) from http://java.sun.com/products/archive/

Run the installer. Set up JAVA_HOME and CLASSPATH environment variables after installation.

c. Eclipse

You will need to use version 3.2.1 of the Eclipse SDK (eclipse-SDK-3.2.1-win32.zip), which can be found at http://archive.eclipse.org/eclipse/downloads. If you install Eclipse, be sure to install it in an area separate from any previous Eclipse installations.

Version 1.3 of the *i2b2 Workbench* has been released and it can use both Eclipse 3.2.1 and 3.3.2 versions. Please consult the *i2b2 Workbench* documentation.

To install, extract the zip file into a directory on local disk. Create a local desktop shortcut to eclipse.exe.

d. yEd Graph Editor

The Correlation Analysis Cell uses the yWorks' yED Graph Editor for viewing and editing Relevance Networks graph files. Download the most recent version for your platform available at http://www.yworks.com/en/products_yed_about.html. After installing, make sure the GRAPHML file format is opened by default by yEd.

2

Tutorial Session

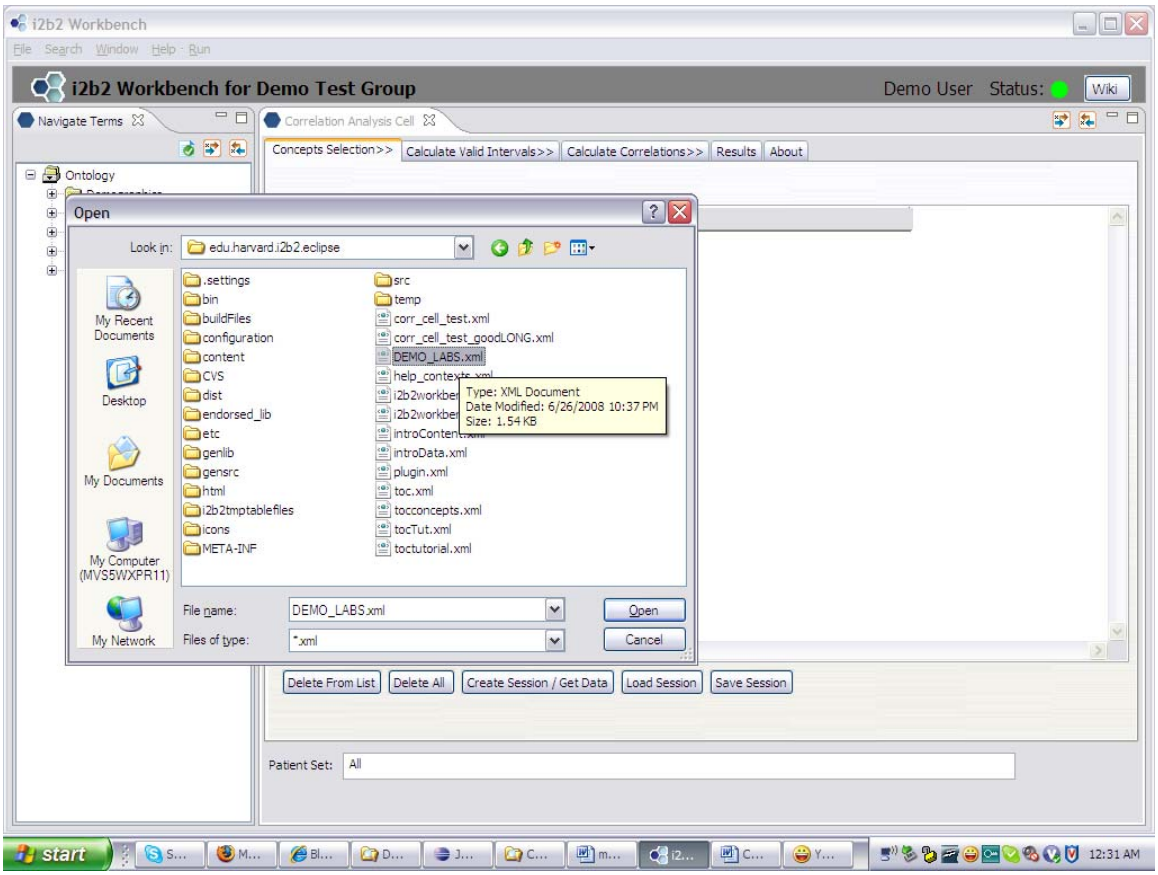
Overview

This *Tutorial* uses an example of correlating three fictitious concepts of type diagnostic labs to guide the user through all computational steps the *Correlation Analysis Cell* algorithm.

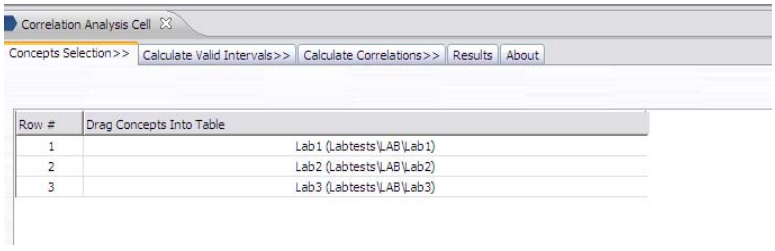
The input data, *valid intervals* and the *normal* correlation computation are presented. Some of these quantities are also manually calculated to allow for a direct comparison with the computer-generated results.

Location of Tutorial Session

This session is located in a file called DEMO_LABS.xml. It will be located in the *i2b2 Workbench* directory. If not there consult the *Correlation Cell Installation Guide* or the *Correlation Cell Developer's Guide* for proper location.



After import one should see the following configuration:



Next click on *Create Session/Get Data* button.

Input Data for all lab types and patients

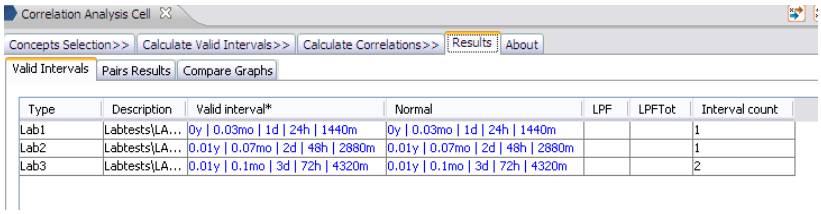
Names of the types are “Lab1”, “Lab2” and “Lab3” (or “1”, “2” and “3” respectively).

We next list the table with the all input lab values, their type, value, the patient they belong to and the time_of_exam parameter:

Type	Result	Patient	Test Date
1	10.00000	pat1	2006-04-25 17:10:24.000
2	20.00000	pat1	2006-04-25 17:10:24.000
3	30.00000	pat1	2006-04-25 17:10:24.000
1	11.00000	pat1	2006-04-26 17:10:24.000
2	21.00000	pat1	2006-04-27 17:10:24.000
3	31.00000	pat1	2006-04-28 17:10:24.000
3	300.00000	pat2	2006-05-01 17:10:24.000
2	200.00000	pat2	2006-05-02 17:10:24.000
1	100.00000	pat2	2006-05-03 17:10:24.000
3	301.00000	pat2	2006-05-10 17:10:24.000

Calculation using the *Correlation Analysis Cell*

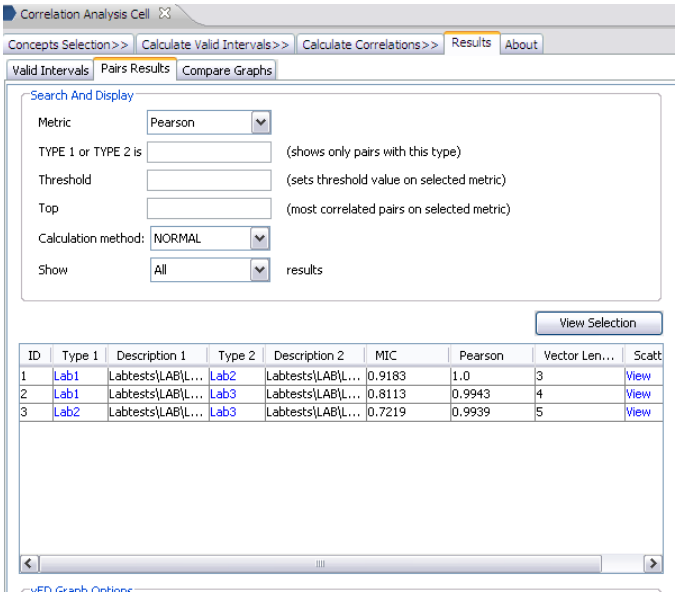
Run the *valid interval* portion of the analysis. Use *normal* method with the default parameter selections.



Go to *Results->Valid Intervals* to display the results for this phase.

Valid intervals (in minutes), *normal* finder are:

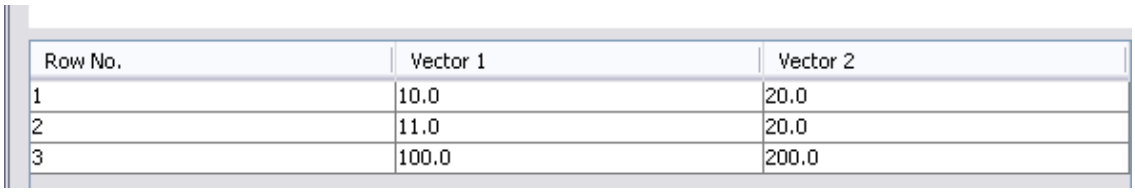
Lab type	Valid interval	
1	1440	(1day)
2	2880	(2 days)
3	4320	(3 days)



Next, go to calculate correlations between the 3 concepts. There are 3 possible pairs between these concepts. Use *normal* overlap method, accept all other defaults. Go to *Results->Pairs* to review.

The correlation results are shown on the screen shot on the left.

Click *View* on the first pair (ID =1), with Pearson's equal to 1 – the overlapped in time vectors are listed below the scatter plot as:



Row No.	Vector 1	Vector 2
1	10.0	20.0
2	11.0	20.0
3	100.0	200.0

Manual Calculation of *valid intervals* and *Pearson's*: Pair with ID = 1

First, construct the time overlap table for the *normal* method. Consult corresponding chapters in *Correlation Analysis Cell User's Guide*.

There are 2 patients ("patient 1" and "patient 2") that simultaneously have data of type "Lab 1" and "Lab 2". To construct the overlapped vectors one needs to find the amount of common overlap between the data types in the pair, per each patient.

Patient 1:

Test Date	"1" result	"2" result
2006-04-25 17:10:24	10	20
2006-04-26 17:10:24	11	
2006-04-27 17:10:24		21

The *valid interval* for lab type "1" is 1,440 minutes (1 day) and 2,880 minutes for lab type "2".

The resulting vectors from the common overlap from data for "Patient 1" will be

Vector 1 = 10, 11
Vector 2 = 20, 20

Patient 2

Test Date	"1" result	"2" result
2006-05-02 17:10:24		200
2006-05-03 17:10:24	100	

This will add to the previous vector a single component of 100 for lab type "1" and 200 for lab type "2" respectively, because the "2" valid interval covers the date of "1" results.

The vectors for the pair will look like

Vector for lab type "1" = (10, 11, 100)
Vector for lab type "2" = (20, 20, 200).

This result matches the one the *Correlation Analysis Cell* displays in tabular format for the *Scatter Plot* option, see screen shot above.

Pearson's **Coefficient Manual Calculation**

The linear correlation coefficient is calculated as follows:

X mean = (10 + 11 + 100)/3 = 40.3333
Y mean = (20 + 20 + 200)/3 = 80
SUM1 = (10 – 40.3333)(20 - 80) + (11 – 40.3333)(20 - 80) + (100 – 40.3333)(200 - 80) = (-30.3333)*(-60) + (-29.3333)*(-60) + 60.3333*120 = 1819.998 + 1759.998 + 7239.996 = 10819.992

SUM2 = $(-30.3333)^2$ + $(-29.3333)^2$+ 60.3333^2 = 920.10908889 + 860.44248889 + 3640.10708889 = 5420.65866667

SUM3 = $(-60)^2$ + $(-60)^2$ + 120^2 = 3600 + 3600 + 14400 = 21600

RESULT = SUM1 / SQRT(SUM2)*SQRT(SUM3) = 10819.992 / (73.6251 * 146.9693) = 10819.992 /10820.6294 = 0.9999410939^2 = 0.99988219= 0.999902

The *Correlation Cell* gives one (i.e. *Pearson's* = 1) for this pair. The two numbers are equal within the rounding accuracy of the display grid.