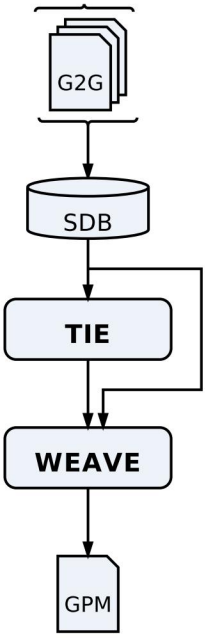
Name: Victor Kofia

TIE & WEAVE outline



G2G — Gene pairscore lists for different tracks.
Formatted as txt files where:
<string:gene> <string:gene> <float:value>

SDB — Systems Database (PostgreSQL)
gene-gene pairs serve as the
primary key

TIE — Integrate pairscore information

WEAVE — Build multigraph from pairscore data

GPM — Gene PairscorGene Pairscore Multigraph
Formatted as txt file where:
<string:gene> <string:gene> <int:track> <float:value>

BCB420: Systemikon TIE & WEAVE outline

Victor Kofia

03/10/2015

Input

Let $N = |LoG|$ where $|LoG|$ is the total # of distinct $< string : gene > < int : accession >$ pairs

INTERACT.txt **# protein-protein interaction pairscores**

1.$< string : gene > < string : gene > < float : value >$
2.$< string : gene > < string : gene > < float : value >$
...
N^2.$< string : gene > < string : gene > < float : value >$

COEX.txt **# co-expression pairscores**

1.$< string : gene > < string : gene > < float : value >$
2.$< string : gene > < string : gene > < float : value >$
...
N^2.$< string : gene > < string : gene > < float : value >$

PATHNE.txt **# pathway neighbour pairscores**

1.$< string : gene > < string : gene > < float : value >$
2.$< string : gene > < string : gene > < float : value >$
...
N^2.$< string : gene > < string : gene > < float : value >$

GOSEM.txt **# gene-ontology semantic similarity based**

1.$< string : gene > < string : gene > < float : value >$
2.$< string : gene > < string : gene > < float : value >$
...
N^2.$< string : gene > < string : gene > < float : value >$

Database Access

N/A

Algorithm

TIE

For database, use open source RDBMS: *PostgreSQL*.

Add pairscore information in **INTERACT.txt**, **COEX.txt**, **PATHNE.txt** and **GOSEM.txt** to database.

Database schema for *SDB*, the *Systemikon Database*:

genes(gid, gname, symbol, organism)

gene_pairs(pid, gid1, gid2)

tracks(tid, tname)

analysis(aid, pid, tid)

scores(aid, score)

INTEGRATE-SCORES(INTERACT.txt, COEX.txt, PATHNE.txt, GOSEM.txt) {
    initialize scoring matrix A where $A = (a_{ij}) \in \mathbb{R}^{N\ x\ N}$ and INTERACT.txt $\rightarrow A$
    $\cdots$ COEX.txt $\rightarrow B$ # $\cdots$ same as above
    $\cdots$ PATHNE.txt $\rightarrow C$
    $\cdots$ GOSEM.txt $\rightarrow D$
    initialize scoring matrix T where $T = (t_{ij}) \in \mathbb{R}^{N\ x\ N}$
    initialize txt file MULT.txt
    for i in $[1, 2, \cdots, N]$
        for j in $[1, 2, \cdots, N]$
            $t_{ij} = a_{ij} + b_{ij} + c_{ij} + d_{ij}$ # Concatenate scores
            add $< i >\ < j >\ < 1 >\ < a_{ij>}$ to MULT.txt
            $\cdots < i >\ < j >\ < 2 >\ < b_{ij>}$ to MULT.txt
            $\cdots < i >\ < j >\ < 3 >\ < c_{ij>}$ to MULT.txt
            $\cdots < i >\ < j >\ < 4 >\ < d_{ij>}$ to MULT.txt
            $\cdots < i >\ < j >\ < 5 >\ < t_{ij>}$ to MULT.txt # 5 tracks in total
    return MULT.txt
}

Let MULT.txt = INTEGRATE-SCORES(INTERACT.txt, COEX.txt, PATHNE.txt, GOSEM.txt)

Add pairscore information in **MULT.txt** to database.

Output

SDB: Systemikon Database

MULT.txt # multigraph

1.$< string : gene >\ < string : gene >\ < int : track >\ < float : value >$
2.$< string : gene >\ < string : gene >\ < int : track >\ < float : value >$

...
$4N^2$.$< string : gene > < string : gene > < int : track > < float : value >$

Tests & Verification

Input? Assumed to have been tested. Multi-graph quality is dependent on input.

Algorithm? Simple unit tests (perhaps with the assistance of an external library) should suffice.

Verification? Ensure that line numbers (of each of the text files) correspond to each other. So, for instance if $|LoG| = 10$ then INTERACT.txt, COEX.txt, PATHNE.txt and GOSEM.txt should all have 100 lines and MULT.txt should have 400 lines.

Database testing? This is tricky ...

From Wikipedia:

Blackbox Testing

- Verify incoming data
- Verify outgoing data from query functions

Whitebox Testing

- Validate database tables & database schema
- Check referential integrity