LH Software Design slide for the BILS cancer project update 02 01 2013

AlgorithmName Duplicator
Preferred Implementation in R/BioC RDuplicator?
Managerial goals BILS deliverable
Methodological goals generic functions, classes, and methods for NetMutatorDuplicator
Publications goals
methodological R/BioC package in Bioinformatics journal
research publication on mutation and duplication data in network context:
(a) duplications fixed/selected depending on network topology?
(b) duplications "broaden bottlenecks"? (shared shortest paths).
(c) distributed robustness versus robustness of system parts.
(d) assymetric mutation load?
(e) DUP1mutated given DUP2mutates (Bayesian-like framework?)
(f) compare all network features between paralogs pairwise: degree, BetCen, dyads.

Methodological goals
PHASE 1:
(1) Annotate cellular network with cancer mutated gene lists (`cancerMutationGeneList` from BROAD ***done*** Oncotator), gene duplication (`duplicationWave`, `duplicationEvent`), and drug target data (`CancerDrugTargets`). Reuse F5 env_fantom5_vectors and env_fantom5_base. ***done***
(2) Visualization of duplication, mutation, and drug target attributes (extends Rgraphviz, or RCytoscape).
(3) Adapt R functions previously developed for functional 2RO analysis, in network context, with p-values from node randomization: `nodeDistribution(), dyadicity(), heterophilicity(), dyadicity_p(), subsetDegree(), subsetGraphDegree(), subsetGraphCC(), subsetBetween(), subsetBetweenTop(), subsetBetween001(), subsetGraphBetween(), bringInEdges(), bringOutEdges(), bringCommonUndirectedEdges(), bringCommonOutEdges(), bringCommonInEdges(), bringCommonScrambledEdges(), bringAllUndirectedEdges().`
(4) Develop new analysis methods for flow of information, robustness, assymetric mutation load, etc (extends RBGL, BioNet, igraph).

PHASE 2 (depends on additional funding):
(1) Use actual cancer mutation data instead of `CancerMutationGeneList` (`cancerMutationEvent`?, extends cgdsr).
(2) Talk directly to TreeFam database (need dedicated TreeFam R package, or use biomaRt?)
(3) Correlate prognosis stats from cBio with network duplication and mutation data.
(4) Make an R package, and distribute through CRAN or BioC.

Depends
R (>= 2.10.0), RCytoscape, BioNet, Rgraphviz, graph, RBGL, BioNet, igraph
Related
HTSanalyzeR (subnetwork GSEA), GeneAnswer (GSEA-like?), NCIgraph (pathws), Mulder2012, ReactomePA (pathws), graphite,

<PHASE1.1>
PHASE1.1 classes
graphNEL and igraph are R/BioC representations of graph objects

Generic function annotateAttribute
use method `annotateAttributeDuplication()` if argument type: `duplicationWave` ***done***
`see youngestFromduplicationWave()`
use method `annotateAttributeMutation()` if argument type: `CancerMutationGeneList`
use method `annotateAttributeDrugTarget()` if argument type: `CancerDrugTargets`
use method `annotateAttributeChrLocation()` if argument type: `ChrLocation`
Generic function annotateEdge
use method `annotateEdgeCoexpressionPC()` if argument type: `CoexpressionPC`
use method `annotateEdgeAgeDistance()` if argument type: `duplicationWave (edges annotated by ageDistance in mln years or first relative with preserved labels)`

Imports
```
RCytoscape.initNodeAttribute
RCytoscape.nodeData
```

arguments: (`graphNEL`) network, (`character`) IDtype, (`data.frame`) duplicationWave, (`data.frame`) CancerDrugTargets returns: (`graphNEL`) annotatedNetwork

Prepare R/Bioc Package Duplicator:
R package consists of code (functions/functions registered as classes: /R), data (duplication, mutation, drug targets: /data), documentation (Rd-style with Rdconv: /man). Fantastic distribution possibilities, and could lead to independent publication in Bioinformatics, BMC Bioinformatics or NAR db issue.

PHASE1.1 Challenges
(1) Multiple duplicationAge tags for nodes: implement so that only the youngest tag is kept: **`*done*`**
(2) Does anything needs to be implemented as S3 or S4 objects?
PHASE2. Challenges
(1) Package techniques: NAMESPACE, R, src, lazyload, data, .Library, library.dynam,