LH analysis flow for the BILS cancer project / BMC Biology paper2 10 01 2013 (week2)
Adapting F5-like environment pipeline for network / cancer mutation pipeline

font conventions used here (following Addison Wesley technical writing style):
`code (Courier New)`, software package/file (Arial), documentation (TNR)

(1) New BioC package igraph significantly extends graph/RBGL.

igraph (Gabor Csardi, Harvard, Barabasi Lab?) reimplements both graph representation and analysis methods, in C from scratch. Resulting package is faster and broader than graph and RBGL (269 pages of documentation).

New duplicator functionalities implemented this week:
a) whole graph analysis with edge weights as co-expression values (GEA, TODO: F5-WP4, time-courses especially):
`edge.betweenness.community()` returns communities object [`plot(as.dendrogram(ebc))`].

b) analysis of **duplicationWaves** (HUMAN, Catarrhini, Chordata, Bilateria, etc):
`duplicationWaveAnalysisWrapper()` calls `subsetDegree(), subsetGraphBetween(), subsetGraphDegree(), subsetBetween001(), subNetwork(), subsetBetween(), subsetBetweenTop(), subsetGraph(), subsetGraphCC(), subsetTransl(), dyadicity(), heterophilicity(), nodeDistribution()`, and returns list with all result vectors.

c) analysis of **duplicationPairs:**
`duplicationPairsAnalysisWrapper` calls `shortest.paths(), betweenness(), degree(), eccentricity(), similarity.jaccard(), similarity.dice(), similarity.invlogweighted()`, and returns list with all result vectors
TODO: add `graph.knn`, and `shortest.paths` with edge weights as co-expression, `edge.connectivity`, `edge.disjoint.paths`, `graph.adhesion`.

d) helper functions:
`getNodesWithAttribute()`
return all nodes annotated with given attribute (TODO: graphNEL->igraph?)
`initializeLayoutWithTaxaShapeColor`
adds node attributes color and shape according to duplicationAge (TODO: graphNEL->igraph?)

e) **env_duplicator_make.R** **main script making the environment with all calculations:**

1. load dependencies and prepare new datasets
2. `annotateEdgesWithPC()` calculate co-expression for connected node pairs & annotate edge weights
`duplicationWaveAnalysisWrapper()` duplicationWave analysis with each taxon separate or grouped into four sets
3. `duplicationPairsAnalysisWrapper()` pairwise duplication comparison, preserve taxon info,
4. save all results in a new environment **env_duplicator**

f) Data: cancer mutation data (duplicator/data/cancerMutationLists/env_cancerMutationLists/); expression data (duplicator/data/expression/) allaffypem_ll (GEA PNAS paper) and F5-CAGE. Network data (duplicator/data/networks/hcsm). Gene duplication data: env_fantom5_vectors, env_fantom5_base

(2) Know how to make a basic R/BioC package now: package.skeleton() and Rd files. The simplest R package is just functions/methods (R/), data (data/), and docs (Rd files). Could use Roxygen2 instead of manual Rd edits.

(3) Next week: make charts and analyze results. Design more precise scientific hypotheses about network evolution and perform targeted tests. Have a draft manuscript by end of January.

(4) Long-term: integrate with ENCODE/F5 promoter-level analysis for TF networks (with OS).

(5) To consider: "duplication and mutation aware" reimplementation of several graph algorithms? In particular, reimplementing shortest path and maximum flow could allow to quantify effect of duplications on distributed robustness, and resistance to mutation (ref. Introduction to Algorithms, Cormen et al.). For example, how many genes in the shortest path are duplicates? Do alternative paths have nodes which are gene paralogs? If you knockout one duplicate (cancer mutation), how much longer does the shortest path become?