The qpgraph package
Network inference and eQTL mapping

Robert Castelo
(robert.castelo@upf.edu)

Dept. of Health and Experimental Sciences
Universitat Pompeu Fabra
Barcelona, Spain

BioC European Developers’ Workshop 2012 - Zurich, Switzerland
Brief overview of the qpgraph package

- Entered BioC release cycle on April 2009 providing functionality to infer molecular regulatory networks from gene expression microarray data.

- It relies on graphical model theory and it is therefore similar to some of the packages in http://cran.r-project.org/web/views/gR.html.

- Its methodology is published in two papers (Castelo & Roverato, 2006, 2009) accumulating together 95 citations in Google Scholar. The BioC site reports an average of 4 downloads a day from distinct IPs.

- Most citations come from other methodological papers, but some independent groups have successfully used it in their analysis of microarray data, such as:


  Oresic et al. Metabolome in schizophrenia and other psychotic disorders: a general population-based study. Genome Medicine, 3:19, 2011.
Brief overview of the qpgraph package

- Network inference by conditional (in)dependence: \texttt{qpCItest()}, \texttt{qpAllCItests()}, \texttt{qpNrr()}, \texttt{qpAvgNrr()}.

- Inference from multiple data sets: \texttt{qpGenNrr()}.

- Model-based estimation of partial correlations: \texttt{qpPAC()}.

- Adjustment for fixed and confounding effects via \texttt{fix.Q} argument.

- Missing data treated via complete-case analysis and EM algorithm.

- Assessment of network quality: \texttt{qpPrecisionRecall()}, \texttt{qpFunctionalCoherence()}.

- Exploration of results: \texttt{qpClique()}, \texttt{qpGraphDensity()}, \texttt{qpGraph()}, \texttt{qpTopPairs()}, \texttt{qpPlotNetwork()}.

- Simulation of data from synthetic networks: \texttt{qpRndGraph()}, \texttt{qpG2Sigma()}, \texttt{rmvnorm()} from \textit{mvtnorm}.

- Parallelization with progress reporting via snow-like MPI clusters.
Long-term goal of the qpgraph package

Gene expression pathway

- Degradation
- Translation
- Nuclear export
- Polyadenylation & Splicing
- Transcription

Markers (n observations)

Samples

Gene expression profiling

Phenotype

Physical map of the genome

Transcript abundance coloured by genotype at a QTL

Transcript abundance coloured by genotype at a marker unlinked to a QTL

Significance threshold

Organismal phenotype Disease

Effective position of gene

Positions of QTLs that explain variation in gene 3 transcript abundance

Transcript-abundance phenotype

QTL genotype

Organismal phenotype Disease

Correlation network

Causal network

Robert Castelo (robert.castelo@upf.edu)

The qpgraph package Network inference and eQTL mapping

Network models aim at exploring a combinatorial number of interactions: gene-gene, gene-phenotype, SV-gene, SV-phenotype:

```r
> sv <- 1e7 ## number of structural variants
> ng <- 25000 ## number of genes
> ph <- 100 ## number of phenotypes
> nl <- sv * (ng + ph) + choose(ng + ph, 2) ## number of links to explore
> nl
```

```
[1] 2.51315e+11
```

```r
> sv <- 1e5 ## number of structural variants
> ng <- 6000 ## number of genes
> nl <- sv * (ng + ph) + choose(ng + ph, 2) ## number of links to explore
> nl
```

```
[1] 628601950
```

Gene expression is a high-dimensional multivariate phenotype vector.

When inferring a SV-gene relationship, one wants to adjust for confounding factors and, ideally, the expression of every other gene.
Network approach in qpgraph: limited-order correlations

- **qpgraph** approaches network inference by testing **repeatedly** (e.g., 100 times) for limited-order correlations of order $q < (n - 2)$:

```r
> library(qpgraph)
> suppressMessages(library(GGdata))
> c20 <- getSS("GGdata", "20", renameChrs="chr20")
> sym2id <- revmap(illuminaHumanv1SYMBOL)
> qpCItest(c20, i=sym2id["PTEN"], j="rs17093026")

Conditional independence test for homogeneous mixed data using an exact likelihood ratio test

data: GI_38505204-S and rs17093026 given {}
Lambda = 0.9172, a = 41.5, b = 1.0, n = 86.0, p-value = 0.02769
alternative hypothesis: true Lambda is less than 1

> qpCItest(c20, i=sym2id["PTEN"], j="rs17093026", Q=sym2id["CPNE1"])

Conditional independence test for homogeneous mixed data using an exact likelihood ratio test

data: GI_38505204-S and rs17093026 given {GI_23397697-A}
Lambda = 0.937, a = 41, b = 1, n = 86, p-value = 0.06939
alternative hypothesis: true Lambda is less than 1
**qpgraph** performs *exact* likelihood ratio tests, instead of one degree of freedom chi-squared tests:

- **Their computational cost grows *linearly* in the size of the conditioning sets \( Q \).**
Biological constraints can provide a major speed-up

- Computations between links are independent and can thus be performed in parallel.

- Not all associations make biological sense. We can restrict calculations to explore the space cisSV-gene, tf-gene, RNA-binding-gene, etc.

*qpgraph* currently enable this in a rudimentary way via the `pairup.i` and `pairup.j` arguments:

```r
> qpAllCItests(exprs(c20)[1:5000, ], estimateTime=TRUE)
```

```
 days  hours minutes seconds
       0      7       59      10
```

```r
> qpAllCItests(exprs(c20)[1:5000, ], pairup.i=1:100, pairup.j=1:5000, 
+       estimateTime=TRUE)
```

```
 days  hours minutes seconds
       0       0      18      15
```
A more friendly interface that can rely on functional terms annotated to features.

> nrr <- qpNrr(smlset, ~ TF*gene + hormone_receptor*gene + cisSV*gene + sex + batch, +          q=20)

The idea would be to identify whether a term refers to a function and then blow it into the set of features annotated to that function.

This can be enabled via the featureData slot in ExpressionSet objects using binary vectors of membership to functional classes.

Should functional names (TFs, RNAbinding, cisSV) belong to some standard controlled vocabulary? (e.g., Homo.sapiens, http://www.sequenceontology.org, ..)
Future plans

- Work with more efficient data structures and a representation in disk that avoids large memory footprints.

- Multicore support for parallelism, via parallel/BiocParallel (need abstraction for reporting progress).
Comments & Bugfixes

robert.castelo@upf.edu
@robertclab #qpgraph

(what about a Tweeter feed #biocbugfix at the BioC site?)

Acknowledgments

Bioconductor project
Alberto Roverato, PhD
Università di Bologna, Italy

Inma Tur, MSc
Universitat Pompeu Fabra

Funding
Spanish MINECO grant TIN2011-22826