Managing very large-scale testing procedures with R

VJ Carey

DSC 2014, Bressanone
Task: genetics of gene expression

• $10^6$ features x $10^9$ variants

• Assay technologies allow consideration of associations that are
  – Tissue-specific
  – Condition-specific

• Slightly different from familiar “big data” problem: problem is not ingestion, but egestion and archiving for further use
Interactive statistical analysis very relevant

• QC, sanity checks
• Model criticism and elaboration
• Want good performance at
  – Storage/access to/modeling of voluminous assay data
  – Retrieval/updating of specific results
Data on IMPUTED genetic variants (‘reals’ in $[0,2]$) can be compressed: David Clayton

snpStats: also includes implementation of glm that uses this representation.
Segments of genome x transcriptome are tested and results are stored to ff as scaled short ints.
Deployment on generic cluster of multicore machines

library(BatchJobs)

csplreg6 =
  makeRegistry(id="mar3",
              seed=123, file.dir="mar3f")

batchMap(csplreg6, doCisChunk,
         1:length(configList)
 )

ids = getJobIds(csplreg6)

submitJobs(csplreg6, ids)
<table>
<thead>
<tr>
<th>(1) Create Registry</th>
<th>BatchJobs’ functions</th>
<th>Common functions</th>
<th>BatchExperiments’ functions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>makeRegistry</td>
<td></td>
<td>makeExperimentRegistry</td>
</tr>
</tbody>
</table>

| (2) Define Jobs     |                      |                  | addProblem                  |
|                     | batchMap             | batchMapResults  | addAlgorithm                |
|                     | batchReduce          | batchReduceResults| makeDesign                  |
|                     | batchExpandGrid      |                  | addExperiments              |

| (3) Subset Jobs      |                      | findDone         | findExperiments             |
|                     |                      | findErrors       |                             |
|                     |                      | ...              |                             |

| (4) Submit Jobs      |                      |                  | submitJobs                  |

| (5) Status & Debugging|                      | showStatus       | summarizeExperiments        |
|                      |                      | testJob          |                             |
|                      |                      | showLog          |                             |

| (6) Collect Results  |                      | loadResult[s]    | reduceResults               |
|                      |                      | reduceResults    | reduceResults[AggrType]    |
Estimation of SNP tendency to be associated with trait variation: bigglm on data.table of 2-20 million records – ROC AUCs, apply over a list of formulas
Upshots

• Aims: achieve feasibility, limit use of resources, facilitate model comparison
• Data acquisition, statistical aggregates, results archiving “transparently” chunked and performed asynchronously
• Constraints: didn’t want/need standard data representations (doubles, .Rdata)
Queries

• “External memory algorithms” seem worthwhile even in the presence of huge quantities of RAM
  – a natural aspect of R software design? A prominent documentation/training objective?

• “Triply agnostic” modeling deployments:
  – Data origins (internal vs. external)
  – Data format (assumed vs. improvised/template)
  – Execution plan (selectable parallelism)