# wcd EST Clustering System

Supplementary Data

March 21, 2008

## Compiler settings

For all tools, we used `gcc` as the key compiler. xsact uses Haskell and we used the GHC 6.6.1 for compiling it.

## 1 Full results of the curated data set

Using Ensembl, we randomly selected 34 non-overlapping genes on the mouse chromosome 4, and used BLAST in dbEST to find ESTs that matched, producing in total 2294 ESTs. We created four reference clusterings M60, M100, M150 and M200: in cluster M$x$, all ESTs that matched a particular gene with at least a score of $x$ were clustered together. The performance of `wcd` and PaCE are shown in Table 1.

| | Sensitivity | | JI | |
|---|---|---|---|---|
| | PaCE | `wcd` | PaCE | `wcd` |
| M60 | 0.58 | 0.66 | 0.58 | 0.66 |
| M100 | 0.62 | 0.70 | 0.61 | 0.68 |
| M150 | 0.66 | 0.73 | 0.64 | 0.70 |
| M200 | 0.70 | 0.75 | 0.65 | 0.67 |

Table 1: Quality of clustering curated data sets

## 2 Full results on the artificial data sets

The series DR00, ..., DR03 were produced using ESTsim, each series with a different error model. Each series consists of 16 EST files (roughly 25k sequences each). These files are available with the supplementary data `http://www.bioinf.wits.ac.za/~scott/wcdsupp.html`, which also has the settings for ESTsim that were used to produce them. DR00 is a set with no errors, DR03 has a moderately high error profile. The quality of `wcd` and PaCE

| | Sensitivity | | JI | |
|---|---|---|---|---|
| | PaCE | wcd | PaCE | wcd |
| DR00 | 0.97 | 0.97 | 1.00 | 1.00 |
| DR01 | 0.92 | 0.97 | 0.95 | 1.00 |
| DR02 | 0.72 | 0.97 | 0.74 | 1.00 |
| DR03 | 0.40 | 0.96 | 0.40 | 0.99 |

Table 2: Sensitivity and Jaccard Index clustering four different artificial data sets

can be seen in Table 2. The results for each series are the average for the 16 files. A deeper, sophisticated analysis would be more informative, but here simply averaging does not skew the results as wcd had a higher or equal index on each set.

# 3 Effects of choosing parameters

We made the decision to test each tool using the default parameters of the tools as provided. The same was done to wcd and we made no attempt to tune wcd for this paper. However, we did carry out some experimentation with one parameter each for PaCE and wcd. The sensitivity and Jaccard index are shown. The results in the table below show the performance on the A076941 data set

**Effect of PacE's EndToEndScoreRatioThreshold**

| EES | SE | JI |
|---|---|---|
| 7 | 0.84 | 0.64 |
| 10 | 0.87 | 0.61 |

**Effect of wcd's window length prate ($l$)**

wcd's default is a window length of 100. As can be seen, this may not be optimal on this data set — a window Long of 125 would have a slightly lower sensitivity but a much higher Jaccard Index.

| $l$ | SE | JI |
|---|---|---|
| 100 | 0.93 | 0.46 |
| 125 | 0.91 | 0.61 |
| 150 | 0.86 | 0.71 |
| 200 | 0.77 | 0.71 |

# 4 Full full results

These are the full results of the comparison of wcd and PaCE on the C4 cluster of the Meraka Instates. In this experiment both wcd and PaCE were compiled using mpicc, O3 optimisation.

The underlying gcc compiler was gcc 3.4.6.

| # slaves | 1 | 2 | 3 | 7 | 15 | 31 | 63 | 96 | 127 |
|---|---|---|---|---|---|---|---|---|---|
| Time (s) | 76243 | 36325 | 24656 | 10652 | 4938 | 2583 | 1411 | 1132 | 878 |
| Efficiency | 1.00 | 1.05 | 1.03 | 1.02 | 1.03 | 0.95 | 0.86 | 0.71 | 0.68 |

# 5 Pthreads parallelisation

The table below shows the cost of clustering the Public Cotton data set on an Intel E5345 dual quad-core machine, using different numbers of cores using the Pthreads parallelisation. We get good speed up going up to 4-5 cores, but beyond that the performance decreases. We think the reason for poor scale-up beyond that is L2 cache contention but this needs further experiments.

| Num cores | Time (s) |
|---|---|
| 1 | 169 |
| 2 | 85 |
| 3 | 61 |
| 4 | 48 |
| 5 | 42 |
| 6 | 39 |