LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MEDIZINISCHE FAKULTÄT
INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG BIOMETRIE UND EPIDEMIOLOGIE

Parallelized preprocessing algorithms for high-density oligonucleotide array data

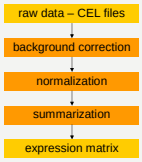
- Markus Schmidberger (schmidb@ibe.med.uni-muenchen.de)
- Ulrich Mansmann (mansmann@ibe.med.uni-muenchen.de)

Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE)
Universität München
Marchioninistr. 15
D-81377 München, Germany
http://ibe.web.med.uni-muenchen.de

Preporecessing

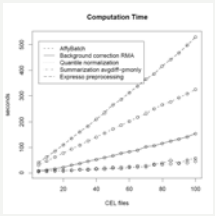
- Background correction
  remove noise of optical detection system
- Normalization
  make measurements comparable from different array hybridizations
- Summarization
  transcripts are represented in multiple probes

- R and Bioconductor mostly used software in research
  - Various different algorithms
  - Some approved and often used methods
    RMA, MAS 5.0, Quantile normalization, Cyclic loess, VSN, expresso, GCRMA



Problems

- Data-structure of R
  - Microarray data are stored in class 'AffyBatch'
  - AffyBatch is memory intensiv
  - 4 GB main memory -> max. 400 CEL files readable (system specific)
- Performance of algorithms
  - Inefficient program structure
  - Long computation time
    Preprocessing for 100 CEL files takes more than 10 minutes (system specific, not linear!)



Challenges

- Microarray experiments more and more popular
- Microarray chips become cheaper
  - Experiments grow in size (6000 Arrays)
- Microarray chips grow in size
  - More genes per chip

Possible Solutions

- Business applications (Expensive, not adaptable)
- Faster and bigger computers (Expensive, limited)
- Better coding (C, DB)
- Distribution to several computers / processors
  - Concurrent calculation of parts at different processors

Parallelization / Distributed Computing

- Multicomputer = Cluster
  different parts of a program run simultaneously on two or more computers that are communicating with each other over a network.

- IBE Cluster
  - 32 workstations
    - 8 GB main memory
    - 2 dual core, Intel Xeon DP 5150
    - 1 Gbit Network
    - LAM/MPI 7.1.3

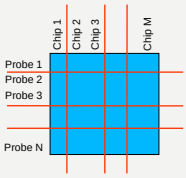
- Message Passing Interface
  - MPI is an API for parallel programming based on the message passing model
  - Libraries exists for: FORTRAN, C, C++
  - R wrapper for the C library available: Rmpi
    - Extensions: Snow, papply

Decomposition

- Data decomposition
  - Data is split in several individual parts and distributed among processes that are essentially similar
  - Task decomposition
    Problem is divided in such a way that each process is doing a different calculation

- AffyBatch stores intensities from multiple arrays
  - Which decomposition is the best ?
    - Partition by chips <-> Partition by probes ?
    - Size of partitions ?

- Partition by chips and partition of CEL File list is a intuitive and good working approach

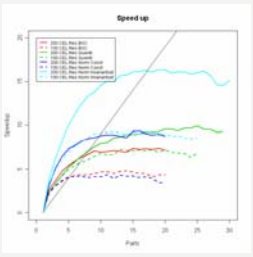


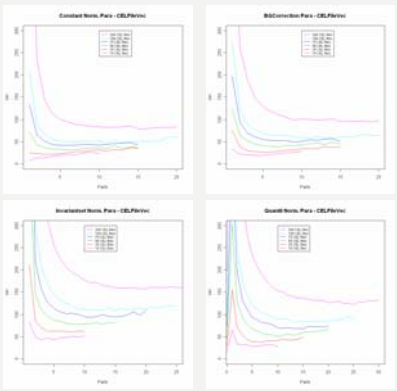
Results

- Package affyPara with parallelized affy-functions
  - More CEL Files preprocessable
  - Speedup
- Considering machine accuracy, parallelized methods produce the same results as serialized methods.

- Partition of data and distribution to several nodes solves the main memory problems
  - IBE Cluster: ~ 16.000 microarrays
- Package is available for testing
  contact schmidb@ibe.med.uni-muenchen.de

- Speedup up to factor 10





Experiences

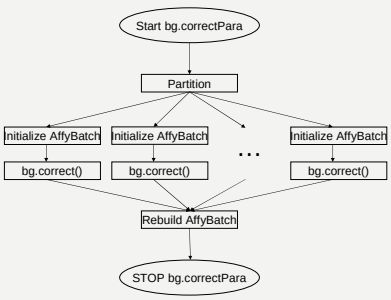
- AffyBatch is memory intensive
  -> a lot of data have to be transferred
- R functions include environments
  -> a lot of data have to be transfered
  -> define global functions
- Use complete Preprocessing algorithms
  -> no AffyBatch to create and to transfer
  -> create small AffyBatches at slaves
  -> CEL files must be available at slaves

Future

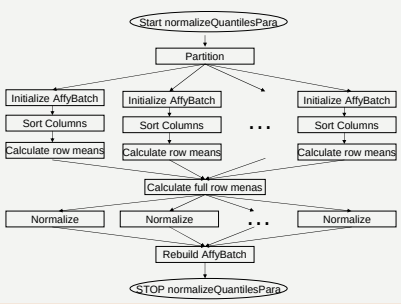
- Rules for ideal decomposition
- Improve summarization (hierarchically)
- Parallelization of other functions: affy, vsn, …
- R and Multiprocessors (openMP)
- R and OpenMPI
- Parallelization of Analyze-Methods

Flowcharts

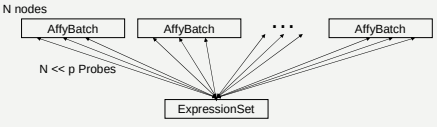
Background Correction



Quantil Normalization



Summarization



November 2007