Data handling strategies for high throughput pyrosequencers

Trombetti GA (1,2), Bonnal RJP (2), Rizzi E (2), De Bellis G (2), Milanesi L (2)

(1) Consorzio Interuniversitario Lombardo per l'Elaborazione Automatica, Milano
(2) Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Milano

Motivation

New high troughput pyrosequencers such as the 454 Life Sciences GS 20 are able to massively parallalize DNA sequencing providing an unprecedented rate of output data and potentially reducing costs. However, these new pyrosequencers also bear a different error profile and provide shorter reads than those of a more traditional Sanger sequencer. These facts pose new challenges regarding how the data are handled and analyzed. More in detail: - Different error profile: new high throughput pyrosequencers provide good performances on average, however, these sequencers have problems in correctly basecalling on homopolymers. - Shorter reads: pyrosequencers nowdays only provide very short reads of about 94 bases on average. This poses problems when a reference sequence is not available or when detecting mutations in repeats or low complexity areas. - Rate: a running 454 LifeSciences GS 20 pyrosequencer can analyze up to 10MBases/hour or 110,000 90-bases-reads/hour. The computation system and data handling strategy should be able to accommodate this.

Methods

To address the challenges described in the above paragraph, we created an automated calculation pipeline integrated with a database storage system. The database is capable of storing, indexing and handling the following information: - Multiple projects of analysis - Biological samples and protocols - Sequences read by the GS 20 sequencer for each run - Multiple co-existing databases of reference sequences - Final results of the calculation pipeline, such as punctual mutations found (with support for heterozygosity) - Intermediate calculations of the calculation pipeline, such as Blast results The pipeline together with the database storage system is capable of easily repeating any past computation thus demonstrating any results obtained, in addition, it is possible to repeat the computations with altered parameters, constants and thresholds and easily compare the results leveraging the database functionalities. The database also allows our biologists to perform inspired researches starting from what they see from the "stock" results obtained from the pipeline. This allows us to discover and investigate peculiar phenomena more easily. The pipeline is multi-stage and mostly parallelizable. Here we quickly outline how we addressed the challenges mentioned in the motivation: Different error profile: Attention is paid in the algorithm so that a nearby homopolymer does not trigger false punctual mutations. Raising the coverage multiplicity only helps up to a point on kinds of errors happening consistently. Shorter reads: The speed and lower operation costs of the 454 machine allowed us to use a high coverage, so that most sporadic errors fall under a threshold. In case of regions similar on more than one reference sequence an heuristic algorithm discriminates which matches are to be kept and which are to be discarded. High data rate: The parallelizability of the pipeline ensures it can keep up with the high data rate of the 454 machine as well as the possibility that biologists would want to repeat a large number of past calculations with different parameters. The database storage system has been set up to store years of operation in an organized and searchable fashion. Backups are made via incremental diffs.

Results

After various runs of the pipeline aimed at tuning the parameters and thresholds for optimal results, we were able to successfully analyze 273 sequenced amplicons from chromosomes 1, 3, 5, 9, 11, 13, 17, 18, 19 of a human sample and correctly find punctual mutations confirmed by either NCBI dbSNP or Sanger resequencing. Sequencing was made with our 454 Life Sciences GS 20 pyrosequencer, obtaining 500,000 reads of 94 bases on average. More analyses will be performed in the future. This pipline is realised in the frame of the Italian MIUR-FIRB project LITBIO www.litbio.org (Laboratory for Interdisciplinary Technologies in Bioinformatics).

Contact email: gabriele.trombetti@itb.cnr.it

Supplementary informations

Trombetti G. A. is a Ph.D student from DEIS department - University of Bologna, Italy