

https://dspace.org/

Parallel algorithms for querying spatial properties in the protein data bank

Selvan, Joshua

2019

https://hdl.handle.net/10539/30383

Selvan, Joshua (2019) Parallel algorithms for querying spatial properties in the protein data bank, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/30383>

Downloaded from a DSpace repository. https://dspace.org/

Abstract

Searching large protein databases for proteins with certain structural properties is expensive. This research explored the use of GPGPUs (General Purpose Graphical Processing Units) in speeding up such structural queries. Brute force and kd-tree spatial data structure algorithms were compared and benchmarked against non-GPU parallel algorithms to assess the effectiveness of using GPGPUs.

This was done with the aim of increasing the speed at which queries against large protein databases can be completed to help mitigate the effect of increasing data set sizes of current protein databases [56].

A set of parallel variations of range search algorithms were developed and implemented in the GPU programming language CUDA and their performances times in completing batch range search jobs were compared against other parallel approach types such as multi-threading and message passing to see if the GPU approaches completed notably faster or slower than more traditional parallelised approaches.

The results showed GPGPUs can construct kd-trees far faster than other parallelised implementations can achieve and that in most scenarios (excluding specific cases such as very low or zero result searches) the GPGPU approaches either matched or performed far better than the other parallelised approaches.

While comparing different GPU algorithms, the complex GPU based kd-tree algorithm performed similarly to a simple GPU brute force range search. This highlighted the benefits of writing code which made the most of the GPU's parallel architecture as opposed to modifying efficient (recursive) algorithms to adequately fit into those same GPU architectures. This implied that even though spatial data structures are effective ways of dealing with protein data, there are better returns on effort in writing code specifically for the GPU's inherently parallel architecture for initiatives which require algorithms to be developed from scratch.