Poster B-66

TraceSearch: Implementation and Applications.

Authors:
Adam Spargo (*WTSI*)
Steve Leonard (*WTSI*)
Guy Coates (*WTSI*)
Jody Clements (*WTSI*)
Roger Pettett (*WTSI*)
Antony Cox (*WTSI*)
Martin Widlake (*WTSI*)
Zemin Ning (*WTSI*)

Short Abstract: TraceSearch enables access to the trace archive via homology. Our distributed implementation of the SSAHA2 algorithm allows data to be added incrementally on a daily basis and hardware failures to be handled efficiently. As well as technical details we present examples of how the system is used in research.

Long Abstract:
The TraceSearch system enables access to all data in the Wellcome Trust Sanger Institute (WTSI) trace archive via sequence similarity. The system has been available as a free web service as of 21st March this year and is steadily gaining users. The main benefits of the system are speed, data coverage and low maintenance.

Due to the effectiveness of the SSAHA2 algorithm for this problem, we are able to search for k-mer word matches over vast amounts of data while performing the computationally intensive alignment phase for only the very best candidates, the traces which show the most such k-mer matches to contiguous regions of the query sequence.

Our distributed implementation of the SSAHA2 algorithm means that we may deploy the system over a cluster of commodity hardware giving a truly scalable and economical solution, with potential for a collaborative GRID implementation to increase reliability. We currently run the system on a cluster of 35 dual-CPU IBM LS20 blades, with alignments performed on a 4-CPU HP DLX85. The system is currently non-redundant with only a single instance of each hashtable in memory at any given time. This means that hardware failures must be handled efficiently. When a node fails, fail-over to a replacement involves the copying of the relevant hashtable files from a remote location and the loading of the data structure into memory before the server can once more accept traffic. This process takes a small number of minutes to accomplish once triggered, we are currently looking into the use of the Linux-HA system to automate the process. A queuing system is already in place on the web server which will hold queries during the fail-over.

The crucial part of the system design is that we do not distribute by species, in fact we make no distinction at all about which data is assigned to which node. This has two major advantages. First, there is no species for which the dataset is deemed too small to warrant a

server, so all traces are searchable. Second, there is no need to rebuild the index for every species upon the arrival of new data, only the currently active node has its index updated incrementally on a daily basis. Unlike other indexing schemes the SSAHA hashtable data structure supports a simple binary addition. So new data can simply be hashed and the resulting hashtable merged with that on the active node. As a result the service is rarely more than a day out of date, while the system administration is minimal.

As well as technical details we present concrete examples of how the system has already been used in sequence production and in molecular biology research at the WTSI. The most common application of the service is for validation. For example we can easily ascertain which species a given sequence aligns to, as well as where it was sequenced, this has already resolved suspected contamination issues in various projects.

Another major application is in finishing, problematic regions may be submitted as query sequences to TraceSearch. The resulting hits can then be used to cover gaps using reads that may be too repetitive to be handled effectively by conventional tools.

Localized assemblies can be performed from any given seed sequence, such as EST data. All reads which align to a query can be downloaded in FASTA or SCF format for further processing, the resulting consensus can give a good representation of a genome in a specific neighborhood; even if the whole genome assembly is in bad shape. Furthermore, syntenic regions in other species are immediately identified. Taking an iterative approach, these localized consensus can be grown in length and depth.

In short the service can increase confidence in results by applying all publicly available data to a problem, rather than just data from any given project, adding the maximum available depth at a given locus. Future features include the possibility to exclude a given species from the search, to search using a protein sequence query, more flexible internet services and online post-search applications.