

Searching DNA for Matches and Close Matches

OVERVIEW

Biotechnology applications involve compute-intensive algorithms and require high-performance computing solutions that:

- process huge amounts of data
- accommodate bit-level searching
- facilitate streamed processing

As biotechnology continues to push the computation envelope ever further, Tarari Processors deliver on the promise of high-performance computing acceleration for researchers, scientists and academicians. Tarari Processors are well suited to the high-performance computing needs of biotechnology.

The Human Genome Project has generated huge databases; searching through them for a particular pattern or sequence can take days, weeks or even



months. Common usage types include:

- pattern-matching inside DNA, looking for hits or close hits
- protein analysis and comparisons

In the rigorous mathematical search, the application slows to a crawl.

While comparing a query pattern with deleted or inserted elements against a target database, a typical application will assign a score based on the closeness of a match. This is known as “sequence alignment.” In the case of genomic research, an application compares a

relatively small number of known query patterns or sequences against billions of samples in the dataset, looking for pattern-matches in other parts of a target genomic structure.

This kind of processing task might be within the capabilities of a host processor; however, the time required to achieve a “fuzzy” result quickly exceeds expectations. A key requirement for the application is to discern a pattern where the pattern-match can be fuzzy and inexact. If the match is close, that may be good enough to have research value. There may be one element of the data that’s wrong or out of sequence, but the finding would still have some value to the researcher because there is immense variety in genomic patterns. The application could be set to report all pattern-matches scoring above a given threshold. The complexity of the algorithm and the processing power required to accomplish it rise dramatically

as the level of fuzziness increases.

For example, an application might be examining long strands of DNA for a close match to a particular combination of genetic characteristics. But in the target dataset, the query characteristics may fall between other features. In the rigorous mathematical search, the application slows to a crawl.

The role of high-performance computing in this application, then, is to process data from the database as fast as possible in a streaming mode, and to perform all of the comparison between the query string and the target patterns away from the host processor.

ISSUES

- ◆ Searches performed on remarkably large sets of data
 - ◆ Host processor not suited to quickly identifying and scoring matches
 - ◆ Useful data contained in time-consuming "fuzzy" matches
-



Tarari, Inc. – The award-winning acceleration company designs and produces Tarari Processors that offload and accelerate compute-intensive algorithms used in network security, Web Services and high performance computing environments.

Tarari's customers include independent software vendors, computer manufacturers and networking manufacturers.

To learn how Tarari can help your applications run at wire speeds, contact Tarari.

FOR MORE INFORMATION

Contact your Tarari, Inc. sales representative, Partner or call Tarari, Inc. at (858) 385-5131 info@tarari.com

Visit our Web site at: www.tarari.com

© Copyright 2003 – Tarari, Inc.

Tarari, Inc.
10908 Technology Place
San Diego, CA
U.S.A.

Tarari and the Tarari logo are trademarks of Tarari, Inc. Other company, product or service names may be trademarks or service marks of others.