

Science.

Technology.

Innovation.



COMPUTATIONAL AND  
INFORMATION SCIENCES DIRECTORATE

*Pacific Northwest National Laboratory has developed a new computational tool that is speeding up our understanding of the machinery of life bringing us a step closer to curing diseases, finding safer ways to clean up the environment, and protecting the country against biological threats.*

# ScalaBLAST

## A Scalable High-Performance Sequence Alignment Tool for Rapid Processing of Sequence Data

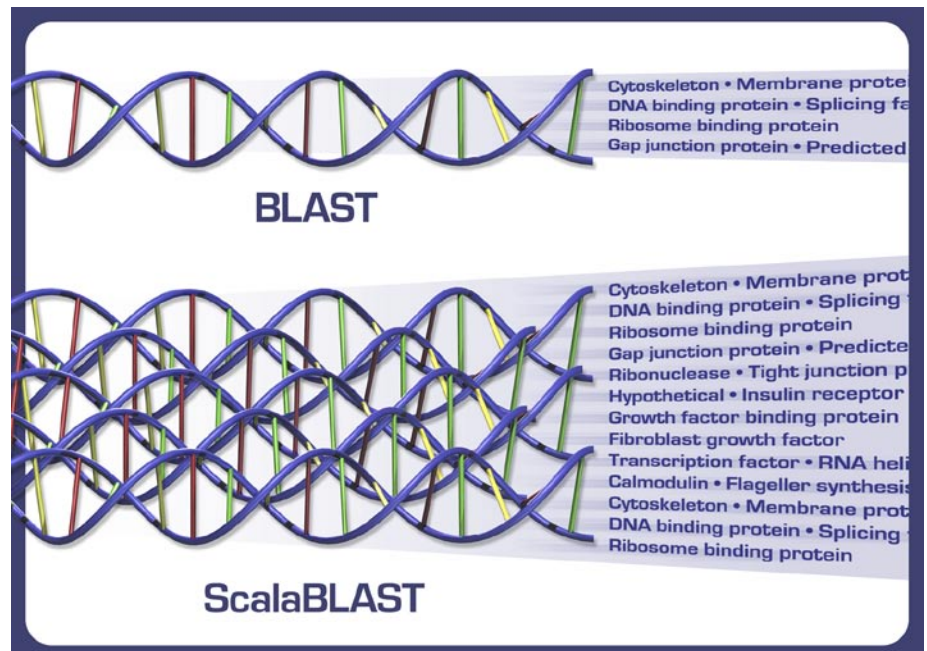
After sequencing the genomes of organisms comes the business of characterizing proteins to understand their functions. Sequence alignments provide a powerful way to compare new sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well-designed queries and alignments.

As public gene and protein sequence databases continue to increase in size at an almost exponential rate, the need to deal with very large database files becomes more urgent. To improve time-to-solution for large-scale sequence alignments, Pacific Northwest National Laboratory (PNNL) researchers scaled up a popular DNA sequence alignment algorithm, BLAST (Basic Local Alignment Search Tool), to become a high-performance, distributed computing alignment tool called ScalaBLAST.

This sophisticated sequence alignment tool can divide the work of analyzing biological data into manageable fragments so that large data sets can run on many processors simultaneously. The scaled-up version of BLAST to ScalaBLAST means that large-scale tasks such as the analysis of an entire organism can be processed in minutes, rather than weeks.

### Speeding up Genomic Sequence Processing

In the world of high-end computing, researchers assemble systems composed of many processors. However, without special modifications, software doesn't run any



ScalaBLAST gives a sizeable performance boost to BLAST, a conventional sequence analysis tool.

**Pacific Northwest  
National Laboratory**

Operated by Battelle for the  
U.S. Department of Energy

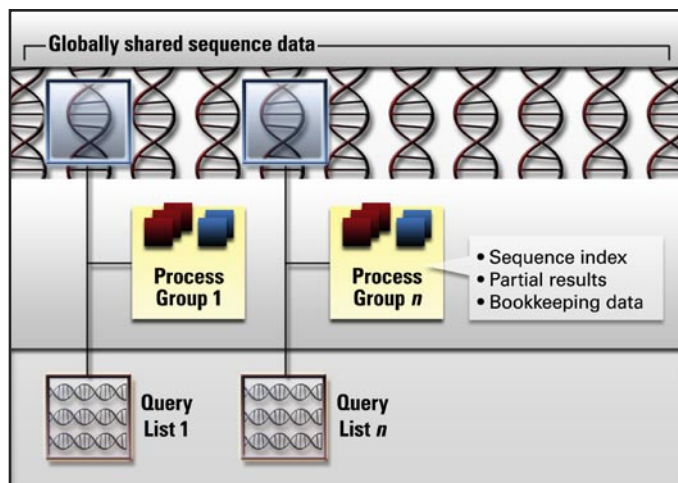


faster on supercomputers than it does on desktop computers. To get answers to complicated biological questions more quickly, researchers “parallelized” the software using Global Arrays, a parallel programming toolkit developed at PNNL, by creating algorithms to divide up the work among the processors. With the BLAST algorithm, it took researchers 10 days to analyze one organism. Using ScalaBLAST, researchers can analyze **13 organisms in about nine hours.**

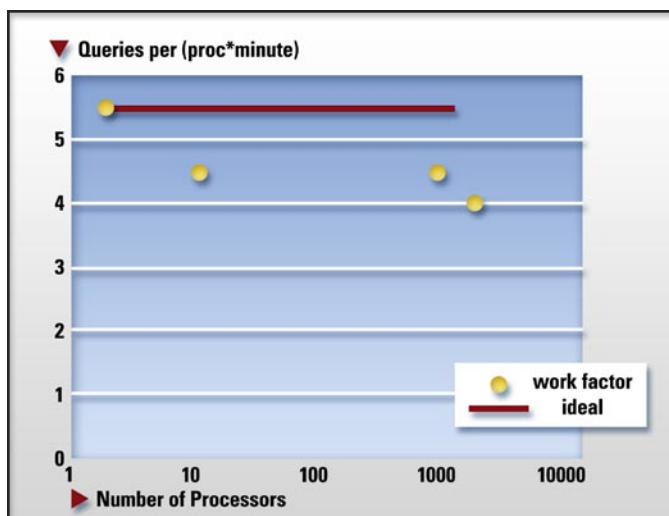
ScalaBLAST was developed using a software-based shared memory approach. The goal was to achieve efficiency and scalability on shared and distributed memory architectures. ScalaBLAST’s design is based on two main concepts: 1) to enable sharing of very large data sets among processors to accommodate growth in the size of publicly available sequence data, and 2) to retain the efficient scaling characteristic of conventional query-scheduling applications.

The scalability of ScalaBLAST is due to a combination of several optimization techniques that are broadly applicable to high performance sequence analysis. These techniques are:

- **Distributing the target database over available memory:** The goal is to eliminate the need for frequent file access during a query and to enable searching against extremely large databases, both of which are expected for informatics-driven science.
- **Multi-level parallelism to exploit concurrency:** Process groups, available through the Global Array toolkit, make it possible to combine the benefits of shared and distributed memory architectures while hiding their disadvantages.
- **Parallel I/O:** Performance is improved by allowing each processor to create its own output file.



Scope and type of data stored in Global Arrays defined on process groups. Each group has its own local copy of the sequence index array, partial results arrays, and all the bookkeeping arrays. The process groups share a single copy of the sequence data array.



Scaling characteristics of ScalaBLAST on MPP2, a distributed memory system. The number of queries being performed by each processor per minute (per million sequences in the target database) stays very close to the ideal even for very large processor counts. This indicates almost perfect scaling.

- **Latency hiding:** Nonblocking operations in the Global Array toolkit make it possible to hide memory latency (time it takes to retrieve data). This makes ScalaBLAST work as well on distributed memory systems as it does on shared memory systems.

ScalaBLAST is a product of PNNL’s Advanced Computing Technology Laboratory, supporting research projects associated with high-end computing. Development of ScalaBLAST was funded primarily by the U.S. Department of Energy’s (DOE’s) Office of Advanced Scientific Computing Research as part of the BioPilot project, a larger joint research effort between PNNL and Oak Ridge National Laboratory.

This research was performed in part using the Molecular Science Computing Facility (MSCF) at the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL), a national scientific user facility sponsored by DOE Office of Science’s Biological and Environmental Research program. EMSL is located at the Pacific Northwest National Laboratory in Richland, Washington.

#### For more information, contact:

T.P. Straatsma  
 Computational Biology and Bioinformatics  
 Pacific Northwest National Laboratory  
 P.O. Box 999  
 Richland, WA 99352  
 E-mail: tps@pnl.gov  
 Telephone: (509) 375-2802