BLAST

How 90,000 lines of code helped spark the bioinformatics explosion | **By Anne Harding**

You've just cloned and sequenced a gene, but you don't know what it does. Now what do you do? In the absence of functional clues, it's hard to know where to start. One approach is to ask what other known sequences are similar to yours, thereby inferring function from homology.

Each weekday, some 200,000 or so researchers do just that, asking a server at the National Center for Biotechnology Information (NCBI) in Bethesda, Md., to compare their particular sequence against GenBank, a DNA database that, at the end of 2004, held more than 40 million sequences totaling 44.5 billion nucleotides. The NCBI devotes 158 two-processor computers to those queries, 75% of which return within 22 seconds.

The software these servers use, a sturdy 15-year-old program known as the Basic Local Alignment Search Tool, or BLAST, remains, for many, bioinformatics' "killer app." It wasn't the first DNA database search tool, but it was fast, and it provided metrics to assess the significance of the matches it found—all in 90,000 lines of C code.

"The fact that every biologist has been using BLAST tells everything," says Jin Billy Li of the Washington University Genome Sequencing Center in St. Louis, who has used BLASTP (a protein homology tool) to identify flagellar genes in several species, including the human gene that causes Bardet-Biedl syndrome, a ciliation disorder.

The program is so pervasive it has become both noun and verb (as in, "I BLASTed my sequence"). Says Alan Christoffels, director of the Computational Biology Group at the Temasek Life Sciences Laboratory of the National University of Singapore, "As one of our bioinformatics personnel put it, 'it's the Google search [engine] of various genomes.'"

EARLY SUCCESS More than a dozen variants of BLAST exist today; all trace their roots back to a 1982 trip to San Francisco, when David Lipman, then a postdoc at the National Institute of Diabetes & Digestive & Kidney Diseases (NIDDK), bumped into UNIX programmer Tim Havell. Havell thought the operating system's search tools could be used as the basis for a program that would search DNA sequences. "It struck me at the time that he was probably right," says Lipman, now director of NCBI. After all, what is DNA but an ordered string of As, Cs, Gs, and Ts?

So Lipman got to work with another NIDDK postdoc, John Wilbur, and the two came up with an algorithm that could search the entire Protein Data Bank of the National Biomedical Research Foundation (NBRF) in less than three minutes, and all eukaryotic sequences in the Los Alamos Nucleic Acid Data Base in under two minutes.

Within months, Mike Waterfield's lab had used the algorithm to identify the similarity between a viral oncogene and the gene for human platelet-derived growth factor, beating another group by a matter of days.[1] Lipman recalls being thrilled by playing a part in a discovery reported on the front pages of *The New York Times* and the *Washington Post,* and wondering whether "serendipitous" matches like this could be expected with any frequency.

Common sense would suggest that they would not. When Lipman, with Bill Pearson of the University of Virginia, developed FASTA, the next step in BLAST's evolution, in 1985, sequence databases were still relatively tiny. FASTA subscribers received one floppy disk containing the program, and a second containing the most current protein sequence database. GenBank contained fewer than 5,000 DNA sequences in April of that year, and was updated on reels of tape and distributed in hardcover books.

Even so, Lipman says, "your chances of finding something useful in doing a search were pretty good." It had nothing to do with human ingenuity, he adds; it's just that the entire tree of life contains only about 1,000 protein families. "That made any search method useful," he says.

THE SEEDS OF BLAST Gene Myers, a computer scientist then at the University of Arizona, began toying with the idea that would ultimately become BLAST back in 1988. He imagined an algorithm that could use a seed string to generate a set of matches that were close to, but not exactly like, the original. "I realized very quickly that that gave me a tremendous amount of sensitivity and much greater speed, much greater filtration efficiency," recalls Myers, now at the University of California, Berkeley.

Myers wanted the algorithm to be deterministic, meaning it was guaranteed to find every string in a database with a certain degree of similarity – say 80% – to another string. Lipman was pushing for a heuristic method. Such algorithms occasionally miss matches, but are also faster than deterministic ones. Lipman ultimately prevailed, and on a visit to the National Institutes of Health, the pair began working together to develop the tool. "It was David who really saw the potential to build something that everybody would use out of the idea," Myers says.

Along with Stephen Altschul, Warren Gish, and Webb Miller, Lipman and Myers built several prototypes and ultimately published BLAST in 1990. (The name was a play on FASTA, as the new program "blasted" by it, returning matches 20 to 30 times faster. "It was smokin'," Myers says.)

Meanwhile, Gish was tweaking the program to run even faster, for example by using compression algorithms and parallel processing. And he freed users from the constraints placed by their own, limited computing power, by writing a client application that would, via the Internet, allow people to put the NCBI's computing power to work for them. "Not only was it the latest database, it was the latest software, and it ran on a very fast eight-processor computer," Gish notes. "People with these crummy little computers that had maybe only 120K of memory would be able to search these multimegabyte databases very fast."

MANY WAYS TO BLAST Yet there remained room for improvement, including the ability to find gapped matches. Though the classical BLAST finds ungapped alignments, regions of similarity sometimes are interspersed with areas of low homology, the position and ordering of which can reveal functional details. Gish, by then at Washington University, released a gapped algorithm called WU-BLAST, in May 1996. "It expands it into something that has potentially more structure in it," he explains. The NCBI published a competing gapped version in September 1997.

That two teams maintain separate versions of the program—NCBI with BLAST, and Gish and his team with WU-BLAST—has helped boost the programs' popularity, says Robert Hubley, a software engineer with the Institute for Systems Biology in Seattle. "The healthy competition factor is part of it, too."

Another variant is PSI-BLAST, an NCBI version that combines statistically significant alignments into a position-specific score matrix that is then used to search the database a second time to increase search sensitivity. Altschul calls PSI-BLAST the Model T Ford of bioinformatics, because it allowed scientists with much less expertise in statistics and computers to search for sequences and understand their significance. "The Model T was easy for everyone to use. This program ran a lot faster and was completely automated. People could use it who would never have been able to use one of these profile methods before," he explains.

Other versions are optimized for high-performance computing. Aaron Darling, now a graduate student at the University of Wisconsin-Madison, helped develop MPI-BLAST, a version optimized for parallel processing, during an internship at Los Alamos. While a single BLADE computer takes 23 hours to query a large portion of the bacterial genome against databases, 128 of these small computers working together do the search in eight minutes, or more than 170 times faster.

MPI-BLAST users include Harvard Medical School and the Lawrence Livermore National Laboratory, which runs the program on its Thunder cluster, the fifth fastest supercomputer in the world. But smaller groups can also use MPI-BLAST to link together eight or 10 PCs.

DIGGING INTO THE ALGORITHM A few factors have helped boost BLAST's popularity, Lipman says. The unity of life at the sequence level meant that even back in the days of databases on floppy disks, meaningful matches could be found. The PC revolution has meant ever faster (and ever cheaper) computers running BLAST could keep pace with the exponential growth in biological data. Finally, the advent of the World Wide Web has made it possible for individual computer users to put the power of NCBI's servers at their fingertips.

Yet these days, thanks to graphical, hyperlinked images, users rarely see BLAST's guts. They can just put in their search string, press a button, and *voila*! Many users don't even bother changing the default settings.

"The potential for mindlessness and to just believe the answer is it and to just go away either satisfied or dissatisfied is certainly there," says Cheryl Kerfeld, director of University of California, Los Angeles' Undergraduate Genomics Research Initiative (UGRI). "Any answer that you get back from BLAST is really just a hypothesis, whereas students tend to regard that as true."

Students in the UGRI are sequencing the genome of the eubacterium, *Ammonifex degensii*, and performing a preliminary annotation of its genome using BLAST. In a weekly lecture, Kerfeld delves into the algorithm, showing students how BLAST actually works and using this information to illustrate underlying principles of biology, chemistry, and evolution. "The ways to explore it to teach and to illustrate concepts are probably one of the underappreciated strengths of it," Kerfeld says.

Kerfeld uses BLAST in her own work characterizing the structure and function of proteins involved in photosynthesis and stress protection in autotrophic organisms. All the knowledge she has gained by learning about which matrices and settings to use in her work has allowed her to exploit BLAST's full potential, Kerfeld adds.

"As one is evolving from a BLAST beginner to a guru, usually what happens is that BLAST becomes more and more complicated," says Washington University's Li. "There are many parameters one can play with; the more one plays with it, the more advantages one may experience."

THE NEXT FRONTIER Despite its popularity, BLAST, in all its many forms, will ultimately become outmoded. But not anytime soon: many researchers stick with BLAST despite newer, more specialized applications that can do certain jobs more efficiently, says Gish. "There's a certain sense of frustration about BLAST in that everyone tries to use BLAST for everything," says Chris

BIOINFORMATICS MILESTONES

1980 World's first nucleotide sequence database, European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Data Library, opens

1982 GenBank established; it has 606 sequences with 680,338 bases

1983 Using Lipman-Wilbur algorithm, Mike Waterfield detects similarity betweeen a viral oncogene and a cellular gene

1985 Lipman and Pearson publish rapid sequence database search tool, FASTA

1986 Applied Biosystems launches first commercial DNA sequencer

Swiss Institute for Bioinformatics (SIB), European Institute for Bioinformatics establish curated protein sequence database, SwissProt

1988 NIH establishes National Center for Biotechnology Information

CLUSTAL, multiple sequence alignment program for proteins and DNA, released

1990 SIB founds Prosite, database of protein families and domains

NCBI team publishes BLAST

Dagdigian, a cofounder of the BioTeam, a Boston-based consulting collective that seeks to bridge the gap between IT and the life sciences.

What the next "killer app" will be is, of course, unclear. Some see a future in microarray data analysis. For Dagdigian, it won't be something sexy like a new scientific algorithm, but rather integration—a whole suite of tools on a scientist's desktop that will link together in an easy-to-use pipeline capable of winnowing the data chaff.

Myers has a different vision. "I personally believe the next wave is going to be microscopy," he says. It's possible to "light up" every single kind of particle involved in the workings of the cell with existing technology, while microscopy makes it possible to watch what's happening. This would mean making microscopy high-throughput, and coming up with computational techniques to interpret the results.

"There's nothing better than just going in and taking a look," he says. ⊗

Anne Harding (aharding@the-scientist.com)

[1] M.D. Waterfield et al., "Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus," *Nature*, 304:35–9, 1983.

Fast Facts

How has BLAST transformed the life sciences: Provided a way to sieve sequence data

When was it developed: 1990

Primary application: Sequence homology searches

Pros: It's fast, and it puts supercomputing power on users' desktops

Cons: It's not the only game in town (though its users evidently think it is)

Key reference: S.F. Altschul et al., "Basic local alignment search tool," *J Mol Biol*, 215:403–10, 1990.

Clinical application: No direct applications; but sequence homology informs gene function, and that in turn informs drug development

The Many Forms of BLAST

BLAST comes in many forms, some of which are listed here. For more information, see www.ncbi.nlm.nih.gov/BLAST/producttable.html

NCBI-BLAST: The original version of the Basic Local Alignment Search Tool

WU-BLAST: Version of BLAST including gapped alignments, developed by Warren Gish at Washington University, who also is one of the authors of the original BLAST

BLAT: Compares transcript sequences to a genomic sequence template

BLASTZ: Compares the mouse genome to the human genome

MEGABLAST: Allows for a quick search of very similar sequences. Discontiguous MegaBLAST does the same for divergent sequences

BLASTN: Nucleotide-nucleotide searches

BLASTP: Protein-protein searches

RPSBLAST: Reverse position-specific BLAST, used to search conserved domain database

BLASTX: Searches for translated query against protein database

TBLASTN: Searches protein query against translated database

TBLASTX: Searches translated query against translated database

PSI-BLAST: Position-specific iterated BLAST. Three times more sensitive than BLAST, used to detect weak similarities

PHI-BLAST: Pattern-hit initiated BLAST

GEOBLAST: Allows search of gene expression data

IgBLAST: Immunoglobulin BLAST

SNP BLAST: Single nucleotide polymorphism BLAST

1991 Oak Ridge National Laboratory team develops gene recognition and assembly Internet link (GRAIL) software package for genome annotation

1995 European Bioinformatics Institute opens in Cambridgeshire, UK

1999 NCBI releases RefSeq, non-redundant set of sequences for major research organisms including DNA, RNA and protein products

2000 EMBL/EBI and Sanger Institute establish Ensembl, an automatic annotation system for selected eukaryotic genomes

2001 International Human Genome Sequencing Consortium and Celera Genomics publish first drafts of the human genome

Genome Bioinformatics Group at UC Santa Cruz releases first genome browser

2003 Human Genome Project completed

HOW IT WORKS

BLAST

It's common knowledge that BLAST's Web interface puts the power of the National Center for Biotechnology Information (NCBI) on users' desktops. But what exactly does that mean?

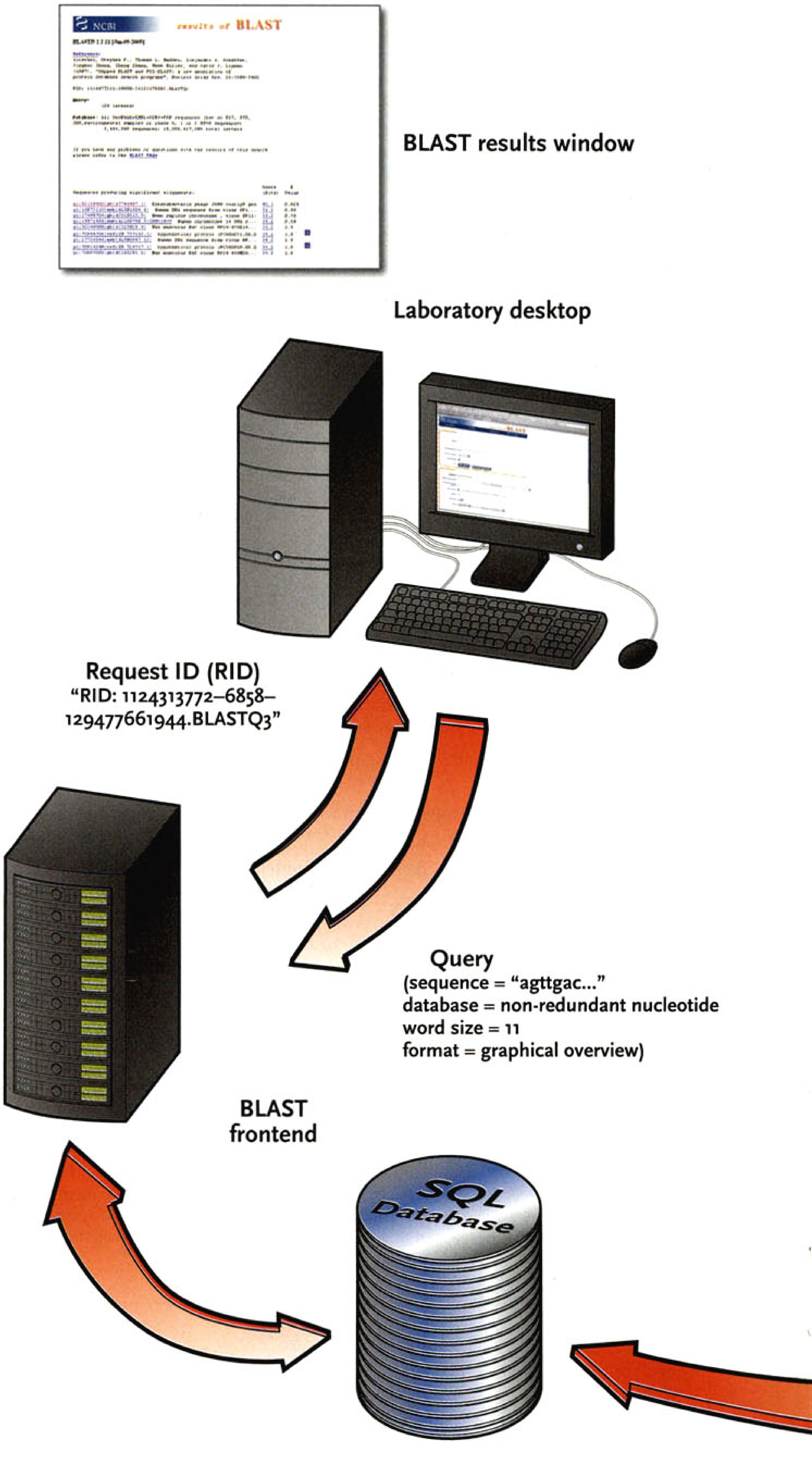
To the user, the process appears seamless: submit a query, wait a few moments, and an answer magically appears. On the backend, however, the process is more complicated. The NCBI receives nearly 200,000 BLAST queries per day, and it devotes some 130 dual-CPU computers to the task of working through those requests.

As NCBI's David Lipman explains, those requests go to the BLAST frontend, which insert them into a structured query language (SQL) database. The SPLITD server daemons then pick up the requests and split them across the BLAST backends ("BLAST Servers"). Separate pieces of the query can thus be processed on up to 10 servers simultaneously.

Most of the sequence data (including GenBank and other sequence databases) is held in a specialized search format on network attached storage devices ("BLAST databases"). But, some of the most-often-used databases are cached locally on the BLAST servers for faster access.

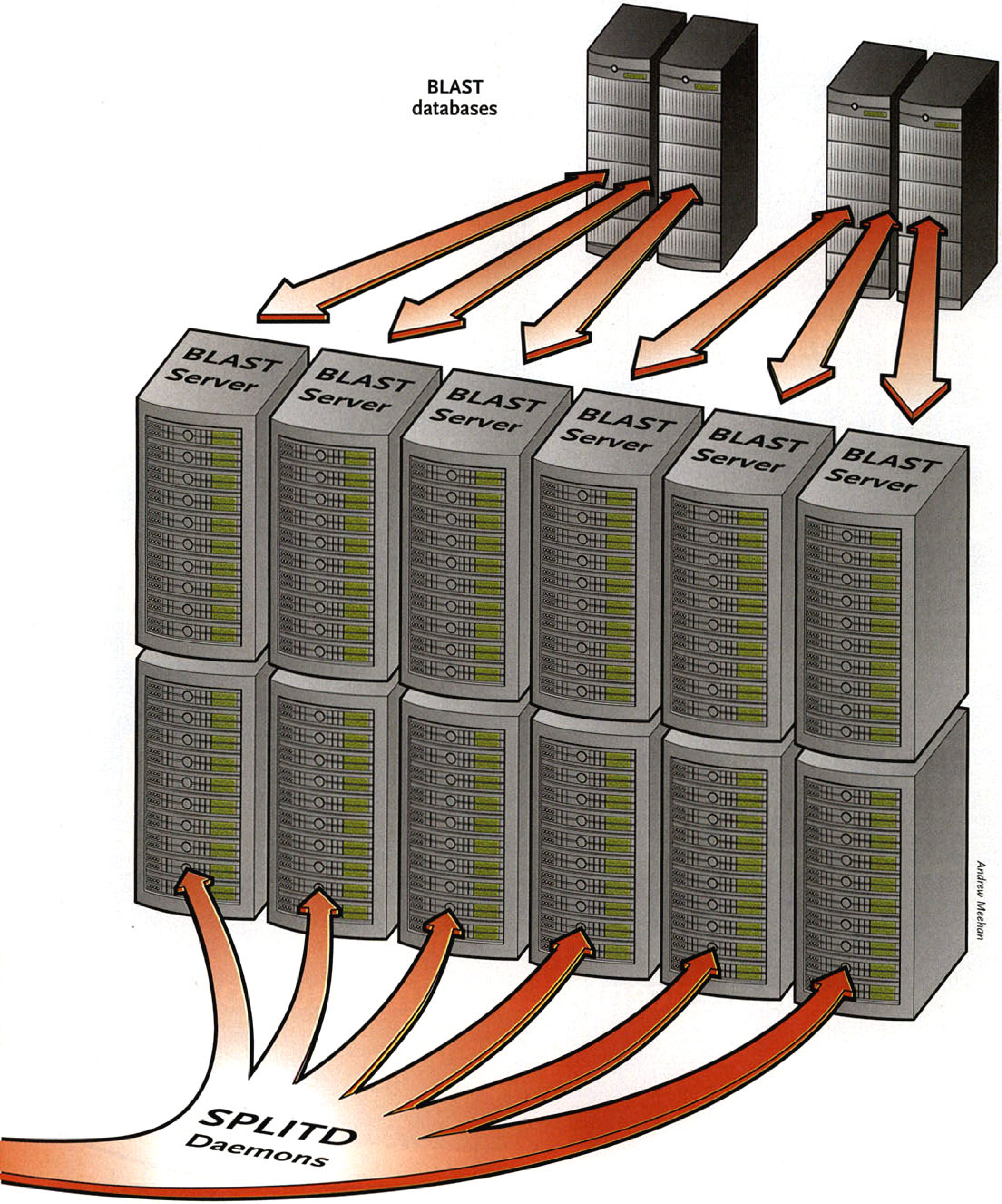
Results from the servers are returned to the SPLITD daemons, which merge the results and store them in the SQL database. When the search is completed, users request the results via the Request ID (RID) assigned to the initial query.

The BLAST frontend assigns a formatting server to fetch the results from the SQL database and return them to the user. Because the alignment results contain no sequence data (rather, they refer to coordinates within the BLAST sequence databases), the server can format them according to the user's specifications (for instance, either as plain text or graphical and hyperlinked) without having to re-execute the search.



"RID: 1124313772–6858–129477661944.BLASTQ3"

(sequence = "agttgac..."
database = non-redundant nucleotide
word size = 11
format = graphical overview)



Andrew Meehan