JCI The Journal of Clinical Investigation

Nucleic acid sequence data turns 100,000,000,000 and looks to the future

Stacie Bloom

J Clin Invest. 2005;115(10):2588-2588. https://doi.org/10.1172/JCI26755.

News

The 3 members of the International Nucleotide Sequence Database Collaboration (INSDC) — the European Molecular Biology Laboratory (EMBL) Bank, GenBank, and the DNA Data Bank of Japan (DDBJ) — have reached a milestone. Owing in large part to their daily exchange policies, these public databases for DNA and RNA sequences have reached 100 gigabases of information. These 100,000,000,000 bases of genetic code, collected since 1982, comprise over 55 million sequence entries from more than 200,000 different organisms. This collaborative effort ensures that information gleaned from molecular biology and genetic research is placed in the public domain where the scientific community can use it to push science forward. “Today’s nucleotide sequence databases allow researchers to share completed genomes, the genetic makeup of entire ecosystems, and sequences associated with patents,” said David Lipman, director of the National Center for Biotechnology Information. “The INSDC has realized the vision of the researchers who initiated the sequence database projects by making the global sharing of nucleotide sequence information possible.” The repositories got started in the 1970s, when researchers suggested a public storehouse be made available for the massive amounts of genetic code sequence information that were being generated. Two of the databases – the EMBL Data Library and GenBank – were launched in the early 1980s. The European Bioinformatics Institute (EBI) manages the EMBL database, […]

years, assuming that it was highly effective.

JCI: What is your typical day like?

Schlegel: I usually get to work around 7 AM so I have about 2 hours to read, write, and think before the rest of the lab arrives. After that, I mainly respond to research and administrative issues. In general, I work after dinner on the internet to screen the literature relevant to our research.

JCI: What do you consider to be your greatest scientific accomplishment?

Schlegel: From a clinical perspective, I think my most significant accomplishment has been contributing to the development of a vaccine that will have a significant positive impact on world health. From a basic science standpoint, I have been most satisfied with our studies of the mechanism by which the E5 oncoprotein alters cell growth by modifying the activity of the cellular V-ATPase.

JCI: What about your greatest life accomplishment?

Schlegel: Overall, my greatest satisfaction is having sufficient time to raise and enjoy a wonderful family.

JCI: By the time this money runs out, what do you hope to have achieved?

Schlegel: Our greatest excitement would come if we could produce an inexpensive, stable HPV vaccine that could be used on patients before *and* after infection.

Stacie Bloom

Nucleic acid sequence data turns 100,000,000,000 and looks to the future

The 3 members of the International Nucleotide Sequence Database Collaboration (INSDC) — the European Molecular Biology Laboratory (EMBL) Bank, GenBank, and the DNA Data Bank of Japan (DDBJ) — have reached a milestone. Owing in large part to their daily exchange policies, these public databases for DNA and RNA sequences have reached 100 gigabases of information.

These 100,000,000,000 bases of genetic code, collected since 1982, comprise over 55 million sequence entries from more than 200,000 different organisms. This collaborative effort ensures that information gleaned from molecular biology and genetic research is placed in the public domain where the scientific community can use it to push science forward.

"Today's nucleotide sequence databases allow researchers to share completed genomes, the genetic makeup of entire ecosystems, and sequences associated with patents," said David Lipman, director of the National Center for Biotechnology Information. "The INSDC has realized the vision of the researchers who initiated the sequence database projects by making the global sharing of nucleotide sequence information possible."

The repositories got started in the 1970s, when researchers suggested a public storehouse be made available for the massive amounts of genetic code sequence information that were being generated. Two of the databases – the EMBL Data Library and GenBank – were launched in the early 1980s. The European Bioinformatics Institute (EBI) manages the EMBL database, while GenBank is the NIH's National Center for Biotechnology Information genetic sequence database.

Both EMBL and GenBank offer an annotated collection of all publicly available DNA sequences, and they were formed as nonprofit entities that collaborated from the beginning. By 1987, the INSDC was formed and included a third collaborator — DDBJ, launched at the National Institute of Genetics in Mishima. DDBJ is also an international nucleotide sequence database freely accessible online.

Early on, staffers searched published journal articles for sequence data and entered it manually into the repository. But times have changed, and the modern sequencing centers at universities today have come a long way. New automated technology, robotics, and bioinformatics, combined with decreased cost, have fostered faster data collection.

"The technology has come so fast that it blows my mind," said Richard Wilson, director of the Genome Sequencing Center at Washington University School of Medicine. Wilson explained that in the mid-1980s his center was able to turn out several hundred bases of sequence per month, while today they are generating about 4.2 billion.

A boost to the number of collected sequences is also due to the National Human Genome Research Institute (NHGRI) at the NIH, which now provides genome-sequencing grants. These funds support research aimed at sequencing a human-sized genome at a cost 100 times lower than is possible now — it presently costs nearly 10 million dollars to sequence the 3 billion base pairs of DNA found in humans. The immediate goal of the NHGRI is to lower the cost of these projects by tens of thousands of dollars in order to allow scientists to sequence genomes of human subjects involved in studies to find genes relevant for disease. The longer-term NHGRI goal is to reduce whole-genome sequencing to only 1,000 dollars so that this process can be used in routine medical tests and allow physicians to tailor diagnosis, prevention, and treatment to a patient's individual genetic makeup.

Large-Scale Sequencing Center at the Whitehead Institute. Today, novel automation technologies and computational packages for genome analysis are available to step up data collection. Photo courtesy of the NIH.

So far, the genetic information available has taught us a lot about the evolutionary relationships among different species. But according to Richard Gibbs, director of the Human Genome Sequencing Center at Baylor College of Medicine, "This will pale into insignificance once we find the full repertoire of alleles that cause different human genetic disease." These are the tools, he explained, that would unlock our understanding of gene function and ultimately provide diagnostic and prognostic indicators. The benefit to human health will be immeasurable.

SB