Pocket K No. 23

Bioinformatics for Plant Biotechnology

Introduction

As of July 30, 2006, scientists around the world are pursuing a total of 2,126 genome projects. There are 405 published complete genomes, and 1,665 ongoing projects. To the field of medicine, this means that there will be a wider field in which to discover potential cures to various diseases. In agriculture, these studies pave the way to understand plant evolution, and use this knowledge to improve crops.

To be able to handle all this genetic information, share and make sense of it, scientists need databases to store the information, where it can be accessed and mined. They also need tools, such as computer software, to manage the information; and algorithms (mathematical formulae) to analyze the information and use it to answer specific questions, such as the location of genes, the structure of proteins, and species relatedness. To do all this (and more), scientists turn to bioinformatics.

What is Bioinformatics?

Bioinformatics is a new science that combines the power of computers, mathematical algorithms, and statistics with concepts in the life sciences to solve biological problems. Through bioinformatics, scientists have been able to analyze various genomes, including those of maize (at http://www.tigr.org/tdb/tgi/plant.shtml) and citrus species (at http://harvest.ucr.edu/).

This Pocket K takes a look at the science of bioinformatics, which can take plant biotechnology from *in vitro* to *in silico,* and where work is moved from the lab to the hard drive (and back to the lab again).

What data does bioinformatics deal with?

Bioinformatics, in general, deals with the following important biological data:

1. DNA, RNA, and protein sequences - The sequence of nucleotides in DNA or RNA, and the sequence of amino acids in a protein, can be obtained through laboratory sequencing methods.

2. Molecular Structures - Higher molecular structure can be obtained by combining thermodynamic data and computer modeling with measurements from laboratory

techniques, such as x-ray diffraction and nuclear magnetic resonance imaging.

3. Expression Data - Scientists use microarrays in the laboratory to determine when and where genes are expressed. Such microarrays can also measure overall gene expressions in certain cell types, or in specific environmental conditions.

4. Bibliographic Data - The number of scientific articles has increased dramatically in the last few decades, due to the increasing number of research projects and genome sequencing programs. These articles are organized in public databases available online.

What can bioinformatics do with this data?
The first step to making sense of all the biological sequences and structures is to formulate a method to manage the data, as well as how to process and maintain it. Data management is the first and most fundamental task of bioinformatics, and bioinformaticians do this by assembling information into databases.

A *database* is a collection of information stored in a systematic way. In bioinformatics, this database may consist of DNA sequences, RNA sequences, or even protein sequences. These sequences may be organized according to their function, or according to the species from which they came, or the journal articles which reported them first. A database may also contain journal articles and abstracts.

With the data assembled, bioinformaticians can find means by which to mine, retrieve, and use the data. This is usually done through computer programs, which can search databases and retrieve information, depending on a scientist's needs.

How can bioinformatics improve plant biotechnology?

•	*It can aid scientists in basic research*

Knowing the complete sequence of a plant's genome can pave the way for all future studies of that organism. For instance, scientists at the United States Department of Agriculture's Agricultural Research Service (USDA-ARS) are now analyzing gene expression patterns in crops such as soybean and barley, in order to determine the function of genes involved in the resistance of plants to environmental stress.

Research teams hail from developed and developing countries alike. The International Rice Research Institute, based in the Philippines, is working on the complete genome of rice. Brazilian scientists have already completed the gene sequence of *Xylella fastidiosa*, a plant pathogen that infects citrus plants.

The worldwide Potato Genome Sequencing Consortium, led by the Netherlands Genomics Initiative and the Wageningen University and Research Center is another example. Teams from countries such as Brazil, Chile, Russia, India, China, Peru, and New Zealand are working together to sequence all 840 million

base pairs of DNA on potato's 12 chromosomes. All this data may be used by scientists to improve potato, which is the world's fourth most important crop.

- ***It can be used to design better plants***

Once the genes responsible for certain plant traits are known, scientists can identify the basis for disease resistance and stress tolerance, and thus design methods by which plants can be made hardier and more resilient. Scientists also use bioinformatics to help them design plants with higher quality fruit, or with the ability to survive in extreme environmental conditions.

Australia's Queensland Agricultural Biotechnology Center, for example, is studying papaya, an important food crop in the tropics, where it is also used in the cosmetics and pharmaceutical industries. To identify the genes involved in papaya ripening, researchers looked at expressed sequence tags (EST) of the fruit's genome. ESTs are short DNA sequences of expressed genes which have been used as a tool for rapid gene discovery. Researchers were able to pinpoint genes that were highly expressed during the ripening process; once these genes are localized, scientists can produce better papayas which may ripen later, or taste better.

- ***It can be used to harness genetic diversity***

By knowing which plants are closely related, scientists can figure out which sexually compatible species have desirable characteristics (such as longer stalks for rice plants, or larger grains for barley, corn, or wheat). The wild relatives of today's plants may be sources of crop improvement genes. Scientists at the University of Wisconsin, for instance, are seeking to improve potatoes by studying the genomes of wild potato species. Researchers at the Weizmann Institute in Israel, on the other hand, are working on understanding the process of gene exchange between crop plants and their wild ancestors, in order to use these processes to incorporate desirable genes from wild relatives into important crop plants.

- ***It can be used to design new tools to study gene function***

Scientists first discovered microRNAs (miRNAs), a family of gene sequences, in plants. These small RNA molecules control various aspects of plant growth and development. They target certain DNA sequences, and, in doing so, keep certain genes from being active. Mutations in miRNAs can cause faulty floral development, or even plant death.

miRNA molecules can be designed to silence whole gene families. As a result, scientists are turning to miRNA technology to develop the next generation of plants. Several projects are now underway in the University of California, Riverside and the Whitehead Institute to predict and identify miRNA families in important crops such as rice.

- ***It can be used to test, analyze, and identify plants***

With more and more microarray profiles online, scientists can learn about and exchange information concerning differences in gene expression. They can also test plants for differences in gene expression or protein profiles under different stress conditions, such as drought, disease,

or insect infestation. If certain genes are expressed in high amounts during these stress conditions, then they may hold the key to a plant's survival under stress - and they may be used to improve other plants that may not have the same gene.

To test if GM plants are comparable to their conventional counterparts, scientists carry out protein or RNA profiling. In a recent research, scientists compared GM potato to conventional potato by analyzing the crops' proteome, and found that there were no new proteins unique to individual GM lines. Scientists from the Danish Institute of Agricultural Sciences used microarrays, as well as analysis software, to compare gene expression profiles of transgenic and wild type wheat. They found that there were no significant differences in gene expression in the two wheat types.

Bioinformatics at your fingertips: The NCBI Online

The National Center for Biotechnology information (NCBI) is an online resource and database for scientists, researchers, and the general public alike. Housed under the United States' National Institutes of Health, the NCBI website is full of tools that can aid interested parties in doing the following:

Search - NCBI contains a search engine called the Basic Local Alignment Search Tool, or BLAST. This search engine is similar to others online, except that the queries are nucleotide (BLASTn) or protein (BLASTp) sequences. Scientists can use the BLAST search to look for DNA or protein sequences similar to those they have. Search matches can then tell them what their gene or protein is, what organism it is from, and what other organisms have the same gene or protein sequence.

Research - ENTREZ is the integrated, text-based search and retrieval system used at NCBI for its major databases. Through ENTREZ, scientists can find out how many genes or proteins of interest are publicly available, how many such genes or proteins have already been sequenced in a given organism, and what research has already been published in the field.

Add - The main database of the NCBI is at GenBank, and sequence "depositors" can add to the nucleotide and protein sequences through an online tool such as BankIt.

Mine - NCBI also has a number of bioinformatics tools available aside from the popular BLAST, all designed to mine data from their online databases. For instance, Spidey can align one or more RNA sequences to a single genomic sequence, and determine where the gene ends and where other sequences begin. If a scientist is working with protein sequences, he/she can use CDArt to see what parts of the sequence are responsible for a given function, and what other proteins have similar domain architectures.

Visit the NCBI at http://ncbi.nlm.nih.gov.

The Way Forward

The more scientists know about plant genomes, the more questions they ask, and the more information they unearth. Bioinformatics not only provides information, but leads to more experiments. For instance, a recent study by scientists from Iowa State University investigated unique sequences in the oat genome. This allowed the researchers to find specific regions of DNA that would both identify oat types through PCR, as well as serve as markers in marker-assisted selection.

Conclusion

There are many tools in bioinformatics, with many functions to suit the needs and expertise of the scientists using them. Gene and protein databases are constantly being updated with information that aid scientists all around the world, in whatever field of the life sciences they are working. Bioinformatics carries benefits for plant researchers: it can aid in plant breeding and genetic engineering, and allow plant scientists to produce better crops for the future.

References
Martienssen, Robert A. Crop Plant Genome Sequence: What Is It Good For? Crop Sci. 44:1898-1899 (2004).
The Crop Biotech Update. http://www.isaaa.org/kc
The Genomes Online Database (GOLD). http://www.genomesonline.org
The National Center for Biotechnology Information. http://ncbi.nlm.nih.gov

Photo Credits: ARS Image Gallery at http://www.ars.usda.gov/is/graphics/photos/.

Pocket Ks are Pockets of Knowledge, packaged information on crop biotechnology products and related issues available at your fingertips. They are produced by the Global Knowledge Center on Crop Biotechnology (http://www.isaaa.org/kc).

For more information, please contact the International Service for the Acquisition of Agri-biotech Applications (ISAAA) SE*Asia*Center c/o IRRI, DAPO Box 7777, Metro Manila, Philippines

Tel: +63 2 845 0563
Fax: +63 2 845 0606
E-mail:knowledge.center@isaaa.org

First printing, August 2006