22 ACEDB: THE ACE DATABASE MANAGER

Jean Thierry-Mieg*, Danielle Thierry-Mieg* and Lincoln Stein**

* Centre National de la Recherche Scientifique, CRBM and Physique Mathématique, 1919 route de Mende 34293, Montpellier, France
** Cold Spring Harbor Laboratory One Bungtown Road Cold Spring Harbor, New York 11777 USA

Overview

The purpose of this paper is to review the main design problems involved in the development of a genome database system and to describe the ace kernel, the stand alone object oriented database manager underlying the C.elegans graphic “Acedb” program.

The ace kernel can handle large amounts of heterogeneous data with a complex evolving schema. It includes a query language and a basic graphic toolbox. It is optimized for speed, memory allocation and disk usage. It has efficient crash recovery and has been intensively tested over the years by demanding biologists. More recently the kernel has acquired client-server capabilities, concurrent write access, visibility over the network, a Java and a Perl interface.

The Ace kernel runs on any Unix workstation or PC with Linux; a port to Microsoft Windows is in the testing phase. It is freely and immediately usable by any interested party. Source code, binaries and documentation can be downloaded from http://alpha.crbm.cnrs-mop.fr

History

Acedb was designed by Richard Durbin and Jean Thierry-Mieg to manage and distribute genetic data on the nematode C.elegans. A survey at the end of 1989 had convinced the authors that an object oriented architecture was desirable but that no existing object oriented manager had the required capabilities to handle the complete C.elegans dataset. A new system was developed from scratch in the C programming
language, a schema was chosen, the data was collated and the graphic C.elegans
Acedb was first released to the worm community in June 91.

Because the source code was made freely available from the start[1], Acedb was
rapidly adopted by communities working on a wide variety of organisms, including
the plant Arabidopsis (Mike Cherry, Sam Cartinhour, then John Morris), the fruitfly
Drosophila (John McCarthy, Suzanna Lewis, Frank Eeckman and Cyrus Hamon), the
yeast S.cerevisiae (Mike Cherry), comparative vertebrates (Jo Dicks and John
Edwards), edible plants (Doug Bigwood and Sam Cartinhour) and man (David
Bentley and also the Integrated Genome Database (IGD) project, with Otto Ritter,
Detlef Wolf, Petr Kocab, Martin Senger, Jaime Priluski). Acedb is also used in
several non-biological applications, including astronomy and semiconductor
manufacturing. A more complete list can be found in the Acedb FAQ (Frequently
Asked Questions list), first assembled by Bradley Sherman in 1994, and now
maintained by Dave Matthews [2].

A number of acedb workshops were organized, successively in Cambridge,
Boston, Montpellier, San Francisco and Cornell. They last a couple of weeks and
allow ample time for discussion and cooperative programming.

At the same time that it was spreading to other organisms, Acedb was becoming
more important to the Genome Project. It has been used by the Sanger StLouis
consortium, led by John Sulston and Bob Waterston, to manage data and annotate
over 100 megabases of DNA sequences and more than 1 million expressed sequence
tags (ESTs). As Acedb grew, we used its rich graphic interface for front ends to a
number of specific analysis programs, allowing trained users to build maps, find
genes (Phil Green) or display multiple alignments (Erik Sonnhammer). Direct live
interfaces to external programs (Blast, the OSP oligos selection program of LaDeana
Hillier, BioMotif of Gerard Mennessier and the Netscape Web browser) were also
made available. A complete system for sequence assembly and editing, called
Acembly, was built over Acedb and is distributed separately at
http://alpha.crbm.cnrs-mopfr/acembly.html

The development over the last few years of the client server architecture,
described below, allows concurrent write access, overcoming the main limitation of
acedb in a multiuser environment.

The fact that so many programmers and end-users were using Acedb had both
good and bad effects. On the one hand, the speed and reliability of the database
manager and of the basic graphic tool box continuously improved, as bugs were
systematically chased and corrected in particular by Simon Kelley and Michel
Potdevin. Detlef Wolf contributed a series of regression tests called "Aquila" to
maintain code quality, and a commercial product called Purify was used to detect
memory leaks. By 1996 the database kernel had reached a very stable state.