A HIERARCHICAL DATABASE DESIGN AND SEARCH METHOD FOR CODIS

J. D. Birdwell [1], **R. D. Horn**[1], **D. J. Icove** [1], **T. W. Wang, P. Yadav** [1], **and S. Niezgoda** [2]
[1]*University of Tennessee*
[2]*FBI Laboratory, Washington DC*

A new search engine for the FBI CODIS DNA database is under development. The search engine utilizes a hierarchical decomposition of the database by identifying clusters of similar DNA profiles and maps to parallel computer architecture, allowing scale up past the current feasible limit of approximately 10^6 DNA profiles. The goal of this project is to support up to 10^8 DNA profiles in the FBI CODIS DNA database while servicing $10^3{}^{-}10^4$ search requests per hour. The search engine is being designed as a back-end component of the existing CODIS system to avoid significant changes to the user interface and modes of use.

Preliminary results on a single processor (Sun Ultra Enterprise 450 server) using a synthetic database of 400,000 profiles, each with allele information at 16 loci, have achieved search time of 5msec or less for exact (high stringency with no equivalent alleles) match requirements, increasing to times of less than 20msec for matches allowing a single missing or unmatched locus and less than 25msec for matches allowing equivalent alleles at one locus. These data points are highly preliminary and are based upon only a few test cases; more extensive testing will be reported at the presentation. This compares with search times of approximately 5sec using the current release of CODIS and a database of 100,000 profiles.

The key benefits of the new search method are logarithmic scale up and parallelization. Logarithmic scale up means that search times increase in proportion to the log of the number of stored profiles. Thus, while increasing the size of the database from 10^5 to 10^8 profiles increases the search time linearly using the existing CODIS software by a factor of 10^3 (for example, from 5 seconds to 1.4 hours), search times are expected to increase using the new method by a factor of $\log_n (10^3)$, for a small integer base n greater than 1. For base 2, the worst case, this increase is approximate a factor of 10.

Parallelization means the database and search method can be divided into N communicating parts, each residing on a separate computer. The search engine design maps to N processors running under PVM (parallel virtual machine; see http://www.epm,ornl,gov/pvm/) software, where N is in the range from m1 to 128. A parallel architecture is proposed using a network of dual processor Intel Pentium III or Xeon computers communicating over an ATM (asynchronous transfer mode) fiber optic network operating at 155Mb/s (OC-3c) and 622Mb/s (OC-12c). This is similar to implementations of the design developed by the Beowolf project (http://www.beowulf.org/). It is anticipated that the implementation will scale by multiples of eight computers (16 processors), and that the database will be segmented into groups, each implemented on a collection of 16 processors. The upper limit of this design appears to be on the order of 10^8 DNA profiles due to memory limitations of the systems and available network bandwidth.