1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910
301/587-8202

Figure 1 is a resolution test chart. It contains several groups of three horizontal and three vertical lines of varying sizes. Each group is accompanied by a numerical value. The values are: 1.0, 1.1, 1.25, 1.4, 1.6, 1.8, 2.0, 2.2, 2.5, 2.8, 3.2, 3.6, 4.0, 4.5, 5.0, 5.6, 6.3, 7.1, 8.0, 9.0, 10.0, 11.2, 12.5, 14.0, 16.0, 18.0, 20.0, 22.5, 25.0, 28.0, 32.0, 36.0, 40.0, 45.0, 50.0, 56.0, 63.0, 71.0, 80.0, 90.0, 100.0, 112.0, 125.0, 140.0, 160.0, 180.0, 200.0, 225.0, 250.0, 280.0, 320.0, 360.0, 400.0, 450.0, 500.0, 560.0, 630.0, 710.0, 800.0, 900.0, 1000.0. The chart is used to determine the resolution of a system by identifying the smallest line pattern that can be clearly distinguished.

MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

1 of 1

Progress Report
DOE Grant no. DE-FG02-93ER61567
Accurate Restoration of DNA Sequences
Gary A. Churchill

Objectives The primary of this project are the development of (1) a general stochastic model for DNA sequencing errors (2) algorithms to restore the original DNA sequence and (3) statistical methods to assess the accuracy of this restoration. A secondary objective is to develop new algorithms for fragment assembly. Initially a stochastic model that assumes errors are independent and uniformly distributed will be developed. Generalizations of the basic model will be developed to account for (1) decay of accuracy along fragments, (2) variable error rates among fragments, (3) sequence dependent errors (e.g. homopolymeric runs). and (4) strand-specific systematic errors (e.g. compressions) The emphasis of this project will be the development of a theoretical basis for determining sequence accuracy. However, new algorithms are proposed and these will be implemented as software (in the C programming language). This software will be tested using real and simulated data. It will be modular in design and will be made available for distribution to the scientific community.

Progress The basic stochastic model for independent, uniformly distributed errors has been developed. It is based on a hidden Markov chain approach that may facilitate some of the proposed generalizations. Using this model as a basis, statistically rigorous algorithms for (1) fragment assembly, (2) consensus sequence estimation and (3) error rate estimation have been developed. The three algorithms together form the basis of a Monte Carlo method for accurate restoration of a sequence given fragment data. These algorithms have been implemented in software for use with test samples of limited size. The algorithms are computationally prohibitive for realistic sized problems. They do however provide a standard by which other (approximate) methods could be judged. Generalizations of the basic method have been outlined but not implemented. This work is described in detail in Churchill (1994) and Churchill and Casella (1993).

A method of computing the probability distribution of a pairwise sequence alignment has been developed (Thorne and Churchill, 1994). Although our emphasis in this work was on studies of sequence evolution, the basic theory and algorithms can be (and have been) adapted to the sequence assembly problem.

A sequence fragment assembly algorithm based on the Monte Carlo method of simulated annealing has been described by Churchill et al. 1993. Software was developed at Los Alamos National Laboratories and is available on request from Christian Burks.

Plans In light of recent developments in base calling technology (see abstracts by Tibbets et al. and Lawrence and Solovyev from Genome Sequencing Conference V) the original objectives of this project will be extended. New methods are available that can produce base calls as probability distributions on the alphabet $\{A, C, G, T, -, *\}$ where $-$ represents undercall and $*$ an overcall. The problem of combining probability distributions on sequences from multiple runs needs to be addressed. I plan to begin work on the problem of combining two such distributions, one obtained in the direct and another in the reverse-complement orientation from a single clone sequence. I also plan to reexamine the base calling methods of Lawrence and Solovyev which appears to be quite promising but needs further studies of its statistical properties. A clear definition of probability distributions on a sequence space will be an essential starting point for any future on probabilistic base calling. This work is ongoing and will be followed wherever it may lead.

I have obtained a large collection of sequence fragments from the Stanford University Yeast Sequencing Project. A preliminary assembly of the fragments has been constructed. This data will be analyzed to develop a prior distribution for position specific error rates. I am also investigating the possibility of using hidden Markov chain methods to develop an objective criterion for "trimming" the ends of fragments data. This a common practice used to clean-up data prior to assembly. It is currently implemented as an ad-hoc algorithm based on the number of ambiguous calls (N) in a window. This work is ongoing and expected completion date is August 1994.

Implementation of the basic restoration algorithm (Churchill 1994) in code suitable for larger sequencing projects will begin in summer of 1994.

Publications

1. Churchill, GA (1994) Accurate Restoration of DNA Sequences. to appear: *Bayesian Statistics in Science and Technology*, ed. R. Kass.
2. Churchill GA and Casella GC (1993) Sampling Based Methods for the Estimation of DNA Sequence Accuracy. Biometrics Unit Technical Report BU-1138-M, Cornell University.

3. Thorne JL and Churchill GA (1994) Estimation and Reliability of Molecular Sequence Alignments. to appear: Biometrics.
4. Churchill GA, Burks C, Eggert M, Engle M and Waterman MS (1993) Assembling DNA Sequence Fragments by Shuffling and Simulated Annealing. Los Alamos National Laboratories Technical Report #LA-UR-93-2287

Presentations

1. Macromolecules, Genes and Computers III, Waterville Valley, NH, August 1993.
2. Bayesian Statistics in Science and Technology: Case Studies, Pittsburgh, PA, October 1993.
3. Genome Sequencing Conference V, Hilton Head, SC, October 1993.

Address for correspondence :

Gary A. Churchill
Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, NY 14853
Tel: (607) 255-5488
Fax: (607) 255-4698

DATE

FILMED

6/15/94

END