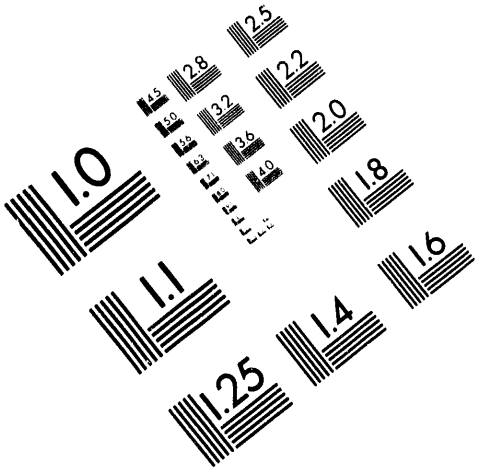


AIIM

Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

301/587-8202

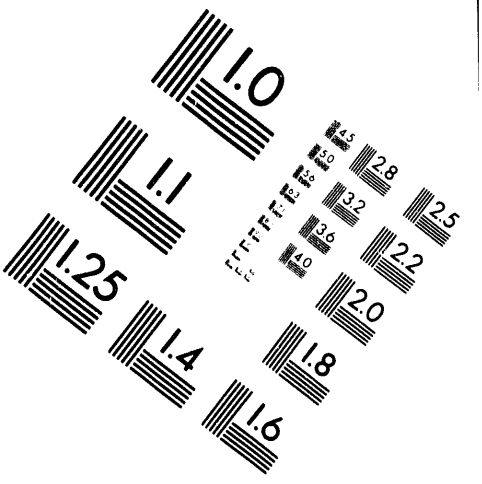
Centimeter

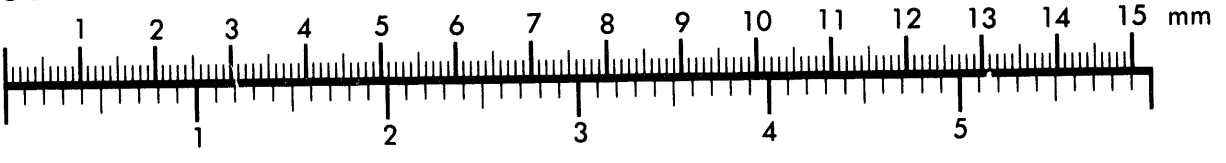
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 mm

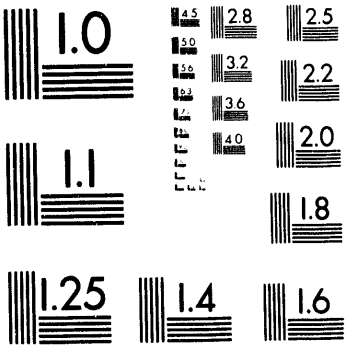
1 2 3 4 5

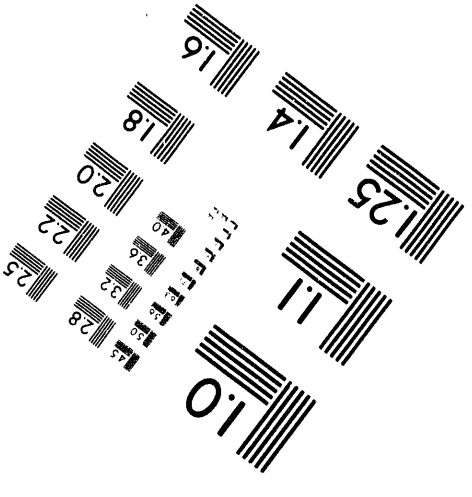
Inches

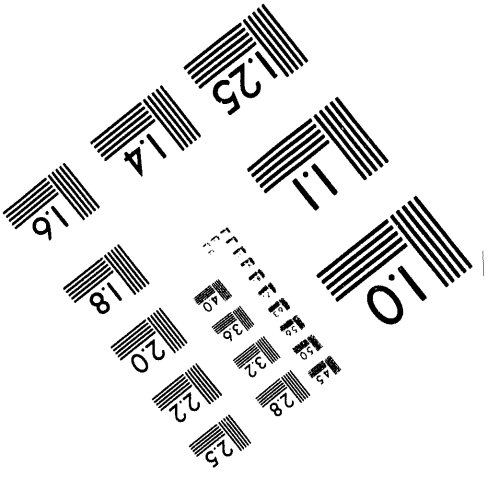
MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.











1 of 1

DOE/ER/61566 - T1

Final Report:

93ER61566

Identification of Genes in Anonymous DNA Sequences

Report Period:
15 April 1993 - 15 April 1994

Christopher A. Fields, PI

The Institute for Genomic Research
932 Clopper Road
Gaithersburg, MD 20878

301-869-9056 (voice) 301-869-9423 (fax)
cfields@tigr.org

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Summary

This Report concludes the DOE Human Genome Program project, "Identification of Genes in Anonymous DNA Sequences," which was supported from August, 1989 through January, 1993 by grant 89ER60865 to New Mexico State University (Final Report submitted 22 April, 1993), and during the present Report Period by grant 93ER61566 to The Institute for Genomic Research (TIGR). The central goals of this project have been 1) understanding the problem of identifying genes in anonymous sequences, and 2) development of tools, primarily the automated identification system **gm**, for identifying genes. The activities supported under the previous award are summarized here to provide a single complete Report on the activities supported as part of the project from its inception to its completion.

Background: Gene Identification and Other Prediction Problems

The identification of genes in anonymous genomic sequences was identified as one of the major outstanding problems in computational biology at the inception of the Human Genome Project (e.g. DeLisi, 1988). The essentially manual, stand-alone sequence analysis tools then available were clearly inadequate for analyzing genomic sequence data at even moderate (10s of kb/day) rates. The **gm** system was developed to meet the need for an integrated, automated gene prediction tool suitable for use on cosmid-sized sequences. Unlike previous tools, **gm** predicted, without user intervention, the complete exon-intron structures, and the amino-acid sequences of the protein products, for all genes contained in a genomic fragment. Following the initial release of **gm** (Fields and Soderlund, 1990), a number of alternative gene prediction systems have been developed, including GeneFinder (P. Green and L. Hillier, unpublished; described in Sulston et al., 1992), GeneID (Guigo et al., 1992), GenLang (Searles, 1992), GAP (Uberbacher et al., 1993), and GeneParser (Snyder and Stormo, 1993). These tools are all capable of assembling exon-intron models of eukaryotic genes, and all have accuracies of about 85% up to 100% of the predicted amino-acid sequence, depending on the organism from which the sequences were obtained, parameter values, and the test sequences used. Accuracies in this range for individual coding exons can now be achieved with neural network methods (Lapedes et al., 1990; Uberbacher and Mural, 1991); most of the predictive power of these systems appears to reside in the measurement of hexamer composition differences between coding exons and noncoding sequences (Bougueleret et al., 1988; Fickett and Tung, 1992).

Once one or more exons are correctly predicted, expression of a gene can be confirmed directly by PCR amplification from cDNA libraries using primers chosen from the predicted exon(s) (Martin-Gallardo et al., 1992; McCombie et al., 1992). Sequencing of such PCR amplification products can be used to rapidly specify gene structure. The availability of reasonably accurate prediction methods and PCR as a follow-up strategy raises the possibility of tightly integrating gene identification with sequencing. Most compositional methods for coding-region prediction, including GRAIL (Uberbacher and Mural, 1991), work well for DNA segments of 100 nucleotides or more, well within the size of typical sequencing runs. Similarity searches against either DNA or protein databases can identify genes efficiently from fragments of this size as well, as demonstrated by the success of EST projects (e.g. Adams et al., 1993). Identifying potential coding sequences as unassembled fragments, and PCR amplifying and sequencing the corresponding cDNAs as part of the sequencing production line, has the potential not only to aid sequence assembly, but also to decrease the need for high average redundancy and hence decrease the cost of genomic sequencing (Fields, 1993). This strategy will be maximally efficient if different classes of noncoding sequences, i.e. introns, untranslated

exons, and intergenic sequences, can be distinguished as well as coding sequences.

Automated gene prediction systems are, however, still neither as reliable nor as informative as similarity searching: the identification of an isologous (significantly similar) sequence in a DNA or protein database remains the most definitive way of identifying a new gene and hypothesizing its function (Seely et al., 1990). Currently, roughly half of identified or predicted protein-coding sequences from all organisms investigated with random genomic or cDNA sequencing approaches are significantly similar, at the protein level, to known genes (summarized in Adams et al., 1993; see also Burland et al., 1993). This observation has led to the suggestion that some sequences have evolved so rapidly that no significant similarities are detectable by alignment methods across phyla (Green et al., 1993). At current sequencing rates, however, large majorities of genes from organisms representing several major phyla, including vertebrates, are expected to be at least partially sequenced as cDNAs within the next few years. The availability of these sequences for searching will allow at least some exons in most, if not all, genes contained in genomic sequences from many organisms to be identified by similarity searching. Only a single member of each isology class identifiable by BLAST (Altschul et al., 1990) or a similar algorithm needs to be sequenced to achieve saturation for gene-identification purposes. Once saturation is achieved, the problem of finding the genes in anonymous sequences will become trivial.

Even if the problem of gene identification can be solved completely, the problems of predicting gene expression and gene-product structure and function remain. Isology provides evidence for structural similarity and functional homology; it contributes little, however, to the prediction of expression patterns. Many genes are proving to be alternatively spliced, alternatively polyadenylated, or expressed under the control of multiple promoters. Complex expression patterns are often correlated with functional plieotropy. The prediction of gene expression patterns, including splicing isoforms, is therefore a major problem for computational genomics. Similarly, an understanding of the histories of gene families, and hence of the degree to which isology can be taken to imply functional homology, is crucial to the elucidation of the functions of newly-identified genes. It is these prediction problems on which longer-term research and development in this area needs to focus.

Project Activities

Overview

The **gm** project was initiated in early 1988 as an effort to automate the identification of genes in genomic DNA sequences. The goal, from the beginning of the project, was to predict the complete exon-intron structures of eukaryotic genes, not just the approximate locations of protein coding regions. The initial **gm** design was described in a poster at the 1988 Genome Mapping and Sequencing meeting at Cold Spring Harbor. A prototype capable of predicting the structures of multi-exon genes in genomic sequences from *C. elegans* with greater than 95% accuracy at the predicted amino-acid level was demonstrated at the Genome Mapping and Sequencing meeting at Cold Spring Harbor, the MacroMolecules, Genes, and Computers meeting at Waterville Valley, and the DOE Contractor-Grantee meeting in Santa Fe in 1989. DOE support for the project under grant # 89ER60865 to New Mexico State University began in August, 1989.

The first distribution version of the software, **gm** 1.0, was released 1 January 1990. **gm** 1.0 was capable of predicting the structures of multi-exon genes in a few seconds on a Unix

workstation. **gm** 1.0 was developed and tested primarily using sequences from *Caenorhabditis elegans* as examples; however, the assembly algorithm applied generally to eukaryotic genes, and users could specify consensus matrices and codon usage tables appropriate to any organism. A substantially faster and more flexible version, **gm** 2.0 was released 31 January 1992. **gm** 2.0 allows the user to specify a wide variety of compositional tests for both exons and introns, and to seed the development of gene models with partial cDNA sequence data. **gm** 2.0 also includes more support for interactive analysis of large genomic regions in its X-windows interface.

Correct prediction of splice sites is the key to accurate prediction of gene structure. At the initiation of the **gm** development effort, the database of known splice sites was small, the mechanism of spliceosome assembly and action was understood only in broad outline, and confirmed cases of alternative spicing were rare. Our initial efforts in sequence analysis, therefore, focussed on assembling better consensus matrices for splice sites and understanding how informative nucleotide positions were distributed around splice sites. An early result of this effort was the observation that the 5' splice sites of the relatively rare long introns in *C. elegans* encode more information than the 5' splice sites of the relatively common short introns. This observation led to a series of comparative studies of intron and exon properties of genes from different phyla.

In July, 1991 the PI began a 1-year visiting appointment in the laboratory of Dr. J. Craig Venter, Section of Receptor Biochemistry and Molecular Biology, National Institute of Neurological Disorders and Stroke. This appointment allowed direct participation in the analysis of two anonymous human 3-cosmid contigs being sequenced by automated methods. **gm** 2.0, GRAIL, BLAST, and other tools were employed in this analysis. A total of 8 genes were identified in a total of 165 kb, including one previously-sequenced gene, two new genes identifiable by sequence similarity to known mammalian genes, and 5 genes not similar to any present in the databases at the time of the analysis. Transcription of the 5 unknown genes was confirmed by PCR amplification of cDNA sequences, using primers chosen from predicted exons. This work demonstrated the utility of an integrated computational and experimental approach to gene identification.

The PI joined The Institute for Genomic Research (TIGR) at its founding in July 1992, and moved the project to TIGR under DOE grant # 93ER61566. The research at TIGR has focussed on characterizing sequence fragments, with an emphasis on developing oligomer-composition tests for predicting coding regions and detecting heterologous sequences in cDNA libraries, and on developing tools for sequence analysis and curation.

gm *and related software development*

The initial **gm** design was based on a two-step method for gene assembly, in which splice sites, translational starts and stops, and polyadenylation sites were identified by consensus matrix methods in the first step, and predicted genes were assembled from the identified sites in the second step. Assembly was carried out from 5' to 3', subject to reading-frame consistency and codon-usage constraints on the candidate coding exons and base composition constraints on both coding exons and introns. The first prototype **gm** software was implemented in 1988, using custom-built modules for sequence manipulation and feature identification and a general-purpose graph-manipulation system (Fields et al., 1988) for representing and assembling the gene models. The sequence analysis modules were implemented in C and ran on a Unix workstation; the graph assembly system ran concurrently on a Lisp machine. This prototype,

which was too slow to be practical for operational use, is described in detail in (Fields et al., 1991).

Work on **gm** 1.0 was started in early 1989. **gm** 1.0 was based on the same method for gene assembly as the initial prototype, with the addition of an initial, low-resolution scan for potential coding sequences, but was implemented entirely in C. It used a set of general-purpose operators originally developed for image analysis applications (Pfeiffer and Soderlund, 1988) for gene assembly. The stringencies of the site identification, base composition, and codon usage tests were set by user-specified parameters. **gm** 1.0 offered two assembly algorithms, an exhaustive algorithm that identified all models consistent with the specified set of analysis parameters, and a greedy algorithm that identified only those models that yielded a maximal protein-coding sequence from a given region of genomic DNA sequence. Both algorithms could be set to identify only models of genes contained completely within the input sequence, or to include models of genes interrupted by one or both ends of the sequence. **gm** 1.0 was tested extensively using genes from *C. elegans*, and regularly achieved prediction accuracies of greater than 95% of the protein coding sequence (Fields and Soderlund, 1990). The program was used to predict several previously-unknown exons of the *unc-22* gene (Benian et al., 1989) and the structures of additional unknown genes contained in the single cosmid sequence (*unc-22*) from *C. elegans* then available. Some of these results are presented in (Schein et al., 1993).

Development of the algorithm, interface, and analysis method enhancements that were incorporated into **gm** 2.0 was initiated soon after the release of **gm** 1.0. **gm** 2.0 uses a substantially more efficient assembly algorithm, which is similar in concept to the "equivalence class" algorithm developed independently by Guigo et al. (1992). Both the initial coding-region scan and the subsequent exon and intron composition tests use a general log-ratio method for comparing the oligomer content, for an oligomer length set by the user, of the test sequences to those of user-supplied standards. A similar test is employed to evaluate the regions immediately up- and downstream of candidate splice sites; use of this optional test improves the false-positive rate for splice-site identification by greater than a factor of two in some sequences. **gm** 2.0 includes facilities for selecting against exons containing known interspersed repeats, and for seeding model generation with partial cDNA (e.g. EST) sequences. The user interface allows the simultaneous display of multiple gene models, positions of interspersed repeats and STSs, and exon positions of known cDNAs, and provides seamless access to a suite of interactive analysis tools. **gm** 2.0 is described in (Soderlund et al., 1992); the software is currently available by anonymous ftp from ftp.tigr.org.

A variety of additional stand-alone software tools have been developed to support the project. These include tools for:

- Parsing the flat-file representation GenBank to extract sequences, annotation, and sequence fragments (e.g. sequences surrounding annotated splice sites).
- Calculating single-position information values and exact uncertainties in these values.
- Counting oligomers in a frame-dependent or frame-independent way, either for whole sequences or using a sliding window.
- Calculating expected oligomer probabilities from observed frequencies of shorter oligomers, using a Markov model.
- Manipulating sequences, generating random sequences for comparative purposes, and calculating standard statistical functions.

These tools are described in (White and Dunning, 1993). They are implemented in Lisp, C, and Unix scripts, and are available by anonymous ftp from ftp.tigr.org.

Splicing signal analysis

The nematode *Caenorhabditis elegans* is a particularly interesting system in which to investigate the mechanism of pre-mRNA splicing, in that it employs *trans*-splicing as well as *cis*-splicing for mRNA maturation (Krause and Hirsh, 1987). As is the case in most organisms studied, the length distribution of *C. elegans* introns is highly asymmetric, with a large number of introns having lengths just above the apparent minimum for successful splicing (Blumenthal and Thomas, 1988). Information distributions for splice sites from short, long, and all available *C. elegans* introns were calculated to better understand which nucleotides contribute to the specificity of splice site selection. Surprisingly, the 5' splice sites of long and short introns were found to have different information contents; the second GU of the canonical GI GUAAGU consensus is conserved in long introns but not in short introns (Fields, 1990). This observation provided the first indication that the mechanism of splice site selection might depend on intron length.

Similar calculations were performed to determine whether evidence for length-dependent mechanisms could be found in other systems. No differences between the splice site structures of long and short introns were identified in either plant or *Drosophila* introns. Dicot plant introns were, however, found to encode less information in their 5' splice sites than monocot introns, but more information in the dinucleotide composition of the intron sequences (White et al., 1992). The principal difference in the 5' splice site structure between monocots and dicots is again in the second GU of the canonical consensus. Differences in 5' splice site structure between long and short introns were also undetectable in *Drosophila*. The positions of putative branch points in *Drosophila* introns are, however, strongly correlated with the positions of the 5' splice sites in short introns, and correlated (as expected) with the positions of 3' splice sites in long introns (Mount et al., 1992). These observations together suggest that the mechanisms of splice site selection may be sensitive to different sequence features in different organisms, and that intron length, in particular, may be important in some organisms but not others, and be measured in different ways in different systems. Experiments in which the addition of as few as 16 nucleotides to a short *Drosophila* intron abolished splicing activity *in vitro* have now provided independent evidence that splice-site selection mechanisms may depend on intron length in at least some systems (Guo et al., 1993).

The role of compositional properties in exon and intron definition, and hence in splice site selection, was further explored by measuring both oligomer composition and deviations from expected composition for oligomers of length k between 1 and 6 for exons and introns separately from *Caenorhabditis, Drosophila,* humans, and both monocot and dicot plants. An order k-1 Markov model was used to calculate expected probabilities. The appropriateness of Markov models as descriptions of DNA sequences is controversial (e.g. Pevzner, 1992). While many groups have used Markov methods to develop principled models of DNA sequences, we use them simply as a phenomenological tool to generate quantitative descriptions of sequences for comparative purposes. Deviations from the Markov-expected probabilities for both exons and introns from each organism or group were measured for each value of k. As expected, stop codons are under-represented in exons, and branch-point sequences are over-represented in introns. In general, however, both exons and introns from the five organisms studied show very complex patterns of deviation from Markov-expected oligomer frequecies, with essentially no detectable similarities between patterns from different organisms for values

of *k* above 3.

The results of the oligomer composition analysis not only confirm the existence of large compositional differences between exons and introns at many values of k, but also show that the compositional properties of both introns and exons vary significantly between taxonomic groups. The latter result suggests that, if compositional properties are detected as part of the splice-site selection mechanism, different compositional properties are likely to be significant in different organisms. This level of complexity is consistent with an emerging view of spliceosome assembly in which different weights are placed on the recognition of different sequence features in different organisms, and even in different introns of the same gene. These complexities will need to be better understood, and taken into account in prediction programs, if accurate prediction of alternative splicing patterns is to be achieved.

Quality control for sequencing projects

All sequencing strategies in which clones to be sequenced are chosen at random from a library are faced with the possibility that some clones may contain DNA from vectors, host cells, source-tissue contaminants, or other heterologous sources. Sequence similarity methods can detect heterologous sequences only if they are already present in the database being searched. The observation of large compositional differences between both exon and intron sequences from different organisms in the work described above suggested that measurement of oligomer composition might provide an assay for detecting heterologous clones in high-throughput sequencing projects. As a test of this idea, a log-likelihood distance measure was used to determine overall differences in hexamer composition between sets of known human, *C. elegans, S. cerevisiae,* and *E. coli* sequences, and between these "control" sets and large sets of human and *C. elegans* ESTs from different libraries.

The hexamer composition test as currently implemented is unable to distinguish sequences from *C. elegans* and *S. cerevisiae,* which appear to have fortuitously similar hexamer composition. All other control distributions tested, however, could be separated at better than 95% confidence by the log-likelihood measure. EST data sets obtained by sequencing randomly-selected clones are expected to include a representative sample of transcribed sequences; they therefore provide an implicit test of the representativeness of the control sets. Both the *C. elegans* (Waterston et al., 1992; McCombie et al., 1992) and human brain (Adams et al., 1991; 1992; 1993a, b) EST sets tested, which represent three different cDNA libraries from each organism, behaved as expected for pure sources of cDNA. However, sequences from one putatively-human EST set, from a commercial lymphoblastoid cDNA library, appeared very similar to the *S. cerevisiae* control. Comparisons of the sequences in this data set (all of which were apparently unidentified when deposited in the EMBL database) to the public databases revealed a large number of exact or near-exact matches to yeast and bacterial sequences, confirming the diagnosis of the hexamer-based test. This result was reported at the DOE Contractor-Grantee and Genome Mapping and Sequencing meetings in Spring, 1993 and in the press (*Science*, 19 Mar 93, p. 1685). A complete description of the method and the tests performed appears in (White et al., 1993), which is included in the Appendix. The test software and the control sets used are available by anonymous ftp from ftp.tigr.org.

Multiple sequence alignment and visualization tools

Conventional multiple sequence alignment algorithms begin the alignment process by computing pairwise alignments, which are then combined to produce a multiple alignment. This process often leads to misplaced gaps, and to spurious expansions of homopolymers. To

address these problems, we have developed a new multiple alignment algorithm, MSA, based on an annealing procedure. The MSA algorithm aligns gap-free blocks of sequence first, and then adds gaps to bring these gap-free blocks into contiguity. MSA is particularly useful for aligning multiple members of protein families that share many relatively small domains separated by weakly-conserved regions of variable length. The MSA package runs on the MasPar SIMD parallel machine, and is callable from the GDE interface developed by Steve Smith. It was first presented at the Genome Sequencing and Analysis Conference in 1993; a paper describing MSA and its applications is in preparation.

A novel feature of MSA is the ability to align relatively dissimilar DNA sequences that encode similar proteins based on an MSA-computed alignment of the protein sequences. This feature allows MSA to compute multiple alignments of DNA sequences too dissimilar to be aligned by conventional algorithms. Protein-driven alignments have proven to be particularly useful in selecting regions of protein-coding genes to use in designing PCR amplification primers that will work across long evolutionary divergences.

The visualization of alignment relations between multiple sequences that are too long to be printed as characters and viewed in an alignment editor is a ubiquitous sequence-management problem. This problem has been addressed with the development of **seqv**, a versatile interface for visualizing multiple sequences, spawning multiple alignments, and editing annotation features. **seqv** was developed for use by sequence data curators engaged in combining and correcting feature annotation for multiple GenBank entries, and includes a back-end interface for passing the derived sequence and annotation information to external programs.

Microbial diversity analysis

Multiple alignments of ribosomal RNA (rRNA) sequences were employed as one of the test cases for the development of the tools described above. These sequences are the main molecular tools used for microbial systematics and taxonomy (Olsen and Woese, 1993; Amann et al., 1994), and many examples are available in public databases. In this course of this work, a surprising degree of variation in rRNA sequences from the same or related species was observed. These differences have been systematically cataloged, and indicate that considerable caution is needed in using rRNA sequences in microbial systematics work, especially in some genera. A paper describing this work is in preparation.

Related activities

Three activities in collaboration with the Center for Human Genome Studies, Los Alamos National Laboratory (LANL), were initiated with partial support from this project. The first of these, an effort to develop restriction map assembly software, was begun in early 1990. This system, CA (Contig Assembler) was based on a fast branch-and-bound algorithm, and allowed a variety of user constraints on the assembly process. It included an X-windows interface similar to that developed for **gm**. A prototype version of CA was installed at LANL in the summer of 1990; CA was described at the DOE Contractor-Grantee meeting in 1991. Work on this problem, using a different assembly algorithm, continued at LANL (Soderlund et al., 1993).

Work was begun on a system for automatically screening sequences of genomic fragments for suitability as STSs in late 1990. This system, ASTS combined identification of known interspersed repeats, similarity searching of known mammalian sequences with **fasta**, detection of potential hairpins, and base compositional tests for the prediction of priming site efficiency. Translations of the genomic fragments were also automatically searched against the protein databases to identify potential coding regions. ASTS was described at the DOE Contractor-

Grantee and Genome Mapping and Sequencing meetings in 1991. Work on this project also continued at LANL (Rappaport et al., 1993).

During the course of the ASTS development effort, a number of sequences were identified as either THE long terminal repeats or as *Mst II* repeats. Comparisons of these sequences to known examples of these repeat classes showed that while the members of both repeat classes are highly diverged from each other, the sequences in both classes are highly similar in several regions, and hence probably share a common ancestor. The new data also demonstrated that these repeats were widely dispersed in the human genome. These results were reported in (Fields et al., 1992).

Personnel.

The principal project staff has included the following individuals. Fellowships and degrees listed were earned in part for work related to this project.

- *Carol Soderlund, Computer Specialist (1989 - 1992), Co-PI (1991 - 1992):* Algorithm and software design and development. Awarded a U. S. DOE Human Genome Distinguished Postdoctoral Fellowship, 1991, to work at Los Alamos National Laboratory. Currently at the Sanger Centre, UK.
- *Pari Shanmugam, Research Assistant (1989 - 1991):* Software development. Awarded M.S. in Electrical Engineering, 1991. Currently an independent software consultant.
- *Owen White, Research Assistant (1990 - 1992), Postdoctoral Fellow (1992 - 1994):* Methods development and comparative sequence analysis, software design and development. Awarded Ph.D. in Molecular Biology, 1992.
- *Granger Sutton, Postdoctoral Fellow (1992 - 1994):* Parallel algorithms, machine learning, multiple sequence alignment.
- *Lisa FitzGerald, Postdoctoral Fellow (1993 - 1994):* Sequence analysis, polymorphisms.
- *Rebecca Clayton, Postdoctoral Fellow (1993 - 1994):* Sequence analysis, microbial systematics.
- *Man Chiu, Software Engineer (1993):* Interface development.
- *Will FitzHugh, Software Engineer (1993 - 1994):* Software and interface development.
- *John Kelley, Systems Administrator (1993 - 1994):* Systems support.
- *Chris Fields, PI:* Software specification and design, sequence analysis, applications.

Five students have participated in the project as part of their training:

- *Ted Slator, Research Assistant (1992):* Software development.
- *Mark DeYong, Research Assistant (Summer 1992):* Neural network applications.
- *Margaret Harloe, Graduate Independent Study Student (Fall, 1989):* Sequence analysis.
- *Linda Rosul, Undergraduate Assistant (1989 - 1990):* Software testing.
- *Chris Short, Undergraduate Assistant (1990 - 1991):* Software development.

Members of the Computing Research Laboratory, NMSU and the TIGR Computational Genomics Facility provided hardware, systems and software support for the project.

Publications

Work supported by this project has been reported in the following papers, copies of which are included in the Appendix.

Fields, C. and C. Soderlund (1990). gm: A practical tool for automated DNA sequence analysis. *Computer Applications in the Biosciences* 6: 263-270.

Fields, C. (1990). Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Research* 18: 1509-1512.

Soderlund, C., P. Shanmugam, O. White, and C. Fields (1992). gm: A tool for exploratory analysis of DNA sequences. In: V. Milutinovic and B. Shriver (Eds) *Proceedings of the 25th Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press. pp. 653-662.

Fields, C., D. Grady, and R. Moyzis (1992). The human THE-LTR(O) and *Mst* II interspersed repeats are subfamilies of a single widely-distributed, highly-variable repeat family. *Genomics* 13: 431-436.

White, O., C. Soderlund, P. Shanmugam, and C. Fields (1992). Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. *Plant Molecular Biology* 19: 1057-1064.

Mount, S., C. Burks, G. Hertz, G. Stormo, O. White, and C. Fields (1992). Splicing signals in *Drosophila*: Intron size, information content, and consensus sequences. *Nucleic Acids Research* 20: 4255-4262.

Fields, C., M. Adams, A. Kerlavage, M. Dubnick, O. White, A. Martin-Gallardo, W. R. McCombie, and J. C. Venter (1992). Identification of genes in genomic and EST sequences. In: H. Lim, J. Fickett, C. Cantor, and R. Robbins (Eds) *Proc. Second Int. Conf. on Bioinformatics, Supercomputing, and Complex Genome Analysis.* New York: World Scientific. pp. 429-433.

Schein, J., M. Marra, G. Benian, C. Fields, and D. Baillie (1993). The use of deficiencies to determine essential gene content in the *let-56 - unc-22* region of *Caenorhabditis elegans. Genome* 36: 1148-1156.

White, O., T. Dunning, G. Sutton, M. Adams, J. C. Venter, and C. Fields (1993). A quality control algorithm for DNA sequencing projects. *Nucleic Acids Research* 21: 3829-3838.

White, O. and T. Dunning (1993). Computational tools for DNA sequence analysis. In: M. Adams, C. Fields, and J. C. Venter (Eds) *Automated DNA Sequencing and Analysis.* London: Academic Press. pp 299-306.

White, O. and W. FitzHugh (1994). A rapid retrieval tool for operating on large flat archive files. Submitted.

Results from the project have been described by the PI or other project staff at the annual Genome Mapping and Sequencing meetings (1988 - 1993), the Genome Sequencing and Analysis conferences (1990 - 1993), the DOE Genome Program Contractor-Grantee meetings (1989 - 1993), the International RNA Processing meetings (1992, 1993), the International *C. elegans* meetings (1991, 1993), and a variety of other meetings, workshops, and departmental seminars.

References.

Adams, M. et al. (1991). Complementary DNA sequencing: Expressed sequence tags and the human genome project. *Science* 252: 1651-1656.

Adams, M., M. Dubnick, A. Kerlavage, R. Moreno, J. Kelley, T. Utterback, J. Nagle, C. Fields, and J. C. Venter (1992). Sequence identification of 2375 human brain genes. *Nature* 355: 632-634.

Adams, M., A. Kerlavage, C. Fields, and J. C. Venter (1993). 3400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genetics* 4: 256-267.

Adams, M. D., M. B. Soares, A. R. Kerlavage, C. Fields, and J. C. Venter (1993). Rapid cDNA sequencing (Expressed Sequence Tags) from a directionally cloned human infant brain cDNA library. *Nature Genetics* (in press).

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.

Amann, R., W. Ludwig, and K.-H. Schleifer (1994). Identification of uncultured bacteria: A challenging task for molecular taxonomists. *ASM News* 60: 360-365.

Benian, G., J. Kiff, N. Neckelman, D. Moerman, and R. Waterston (1989). Sequence of an unusually large protein implicated in regulating myosin activity in *C. elegans*. *Nature* 342: 45-50.

Blumenthal, T. and J. Thomas (1988). Cis and trans splicing in *C. elegans*. *Trends Genet.* 4: 305-308.

Bougueleret, L., Tekaia, F., Sauvaget, I., and Claverie, J.-M. (1988). Objective comparison of exon and intron sequences by means of 2-dimensional data analysis methods. *Nucl. Acids Res.* 16: 1729-1738.

Burland, V., G. Plunkett III, D. Daniels, and F. Blattner (1993). DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: Organizational symmetry around the origin of replication. *Genomics* 16: 551-561.

DeLisi, C. (1988). Computers and molecular biology: Current applications and emerging trends. *Science* 240: 47-52.

Fickett, J. and C.-S. Tung (1992). Assessment of protein coding measures. *Nucl. Acids Res.* 20: 6441-6450.

Fields, C. (1990). Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Research* 18: 1509-1512.

Fields, C. (1993). Integrating computational and experimental methods for gene discovery. In: M. Adams, C. Fields, and J. C. Venter (Eds), *Automated DNA Sequencing and Analysis*. London: Academic Press. pp. 321 - 325.

Fields, C., M. Coombs, and R. Hartley (1988). MGR: An architecture for problem solving in unstructured task environments. In: Z. Ras and L. Saitta (Eds), *Methodologies for Intelligent Systems, 3*. Amsterdam: Elsevier. pp. 40-49.

Fields, C. and Soderlund, C. (1990). gm: A practical tool for automated DNA sequence analysis. *Computer Applications in the Biosciences* 6: 263-270.

Fields, C., H. Pfeiffer, and T. Eskridge (1991). Knowledge representation and control in gm1, an automated DNA sequence analysis system based on the MGR architecture. *Int. J.*

Man-Machine Studies 34: 549-573.

Fields, C., D. Grady, and R. Moyzis (1992). The human THE-LTR(O) and *Mst* II interspersed repeats are subfamilies of a single widely-distributed, highly-variable repeat family. *Genomics* 13: 431-436.

Green, P., D. Lipman, L. Hillier, R. Waterston, D. States, and J.-M. Claverie (1993). Ancient conserved regions in new gene sequences and the protein databases. *Science* 259: 1711-1716.

Guigo, R., Knudsen, S., Drake, N., and Smith, T. (1992). Prediction of gene structure. *J. Mol. Biol.* 226: 141-157.

Guo, M., P. Lo, and S. Mount (1993). Species-specific signals for the splicing of a short *Drosophila* intron *in vitro*. *Mol. Cell. Biol.* 13: 1104-1118.

Krause, M. and D. Hirsh (1988). A *trans*-spliced leader sequence on actin mRNA in *C. elegans*. *Cell* 49: 753-761.

Lapedes, A., C. Barnes, C. Burks, R. Farber, and K. Sirotkin (1990). Application of neural networks and other machine learning algorithms to DNA sequence analysis. In: G. Bell and T. Marr (Eds), *Computers and DNA*. Reading, MA: Addison-Wesley. pp. 157-182.

Martin-Gallardo, A., W. R. McCombie, J. Gocayne, M. FitzGerald, S. Wallace, B. M. Lee, J. Lamerdin, S. Trapp, J. Kelley, L.-I. Liu, M. Dubnick, L. Dow, A. Kerlavage, P. De Jong, A. Carrano, C. Fields, and J. C. Venter (1992). Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genetics* 1: 34-39.

McCombie, W. R., M. D. Adams, J. M. Kelley, M. G. FitzGerald, T. R. Utterback, M. Kahn, M. Dubnick, A. R. Kerlavage, J. C. Venter, and C. Fields (1992). *C. elegans* expressed sequence tags identify gene families and disease gene homologues. *Nature Genetics* 1: 124-131.

McCombie, W. R., A. Martin-Gallardo, J. Gocayne, M. FitzGerald, M. Dubnick, J. Kelley, L. Castilla, L.-I. Liu, S. Wallace, S. Trapp, D. Tagle, L. Waley, S. Cheng, J. Gusella, A.-M. Frischauf, A. Poustka, H. Lehrach, F. Collins, A. Kerlavage, C. Fields, and J. C. Venter (1992). Expressed genes, interspersed repeats, and polymorphisms in cosmids sequenced from 4p16.3. *Nature Genetics* 1: 348-353.

Mount, S., C. Burks, G. Hertz, G. Stormo, O. White, and C. Fields (1992). Splicing signals in *Drosophila*: Intron size, information content, and consensus sequences. *Nucleic Acids Research* 20: 4255-4262.

Olsen, G. and C. Woese (1993). Ribosomal RNA: A key to phylogeny. *FASEB J.* 7: 113-123.

Pevzner, P. (1992). Nucleotide sequences versus Markov models. *Computers Chem.* 16: 103-106.

Pfeiffer, J. and C. Soderlund (1988). Solid modelling data structures for computer vision. *Proc. 22nd Asilomar Conf. on Signals, Systems, and Computers*. New York: IEEE. pp. 543-547.

Rappaport, B., J. Gatewood, C. Fields, and N. Doggett (1993). Integrating repeat identification with thermal calculations. In: M. Adams, C. Fields, and J. C. Venter (Eds), *Automated DNA Sequencing and Analysis*. London: Academic Press. pp. 289 - 293.

Schein, J., M. Marra, G. Benian, C. Fields, and D. Baillie (1993). The use of deficiencies to determine essential gene content in the *let-56 - unc-22* region of *Caenorhabditis elegans*.

Genome 36: 1148 - 1156.

Searles, D. (1992). The computational linguistics of biological sequences. In: L. Hunter (Ed.), *Artificial Intelligence and Molecular Biology*. Cambridge, MA: AAAI Press, pp. 47-120.

Seely, O. Jr., Feng, D.-F., Smith, D. W., Sulzbach, D., and Doolittle, R. F. (1990). Construction of a facsimile data set for large genome sequence analysis. *Genomics* 8: 71-82.

Snyder, E. and Stormo, G. (1993). Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucl. Acids Res.* 21: 607-613.

Soderlund, C., Shanmugam, P., White, O., and Fields, C. (1992). gm: A tool for exploratory analysis of DNA sequences. In: V. Milutinovic and B. Shriver (Eds) *Proceedings of the 25th Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society Press. pp. 653-662.

Soderlund, C., D. Torney, and C. Burks (1993). Calculating shared fragments for the single digest problem. *Proceedings of the 26th Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society Press. pp. 620-629.

Sulston, J. et al. (1992). The *C. elegans* genome sequence project: A beginning. *Nature* 356: 37-41.

Uberbacher, E. and Mural, M. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor neural network approach. *Proc. Natl. Acad. Sci. USA* 88: 11261-11265.

Uberbacher, E., Einstein, J., Guan, X., and Mural, R. (1993). Gene recognition and assembly in the GRAIL system: Progress and challenges. In: H. Lim, J. Fickett, C. Cantor, and R. Robbins (Eds), *Proc. Second Int. Conf. on Bioinformatics, Supercomputing, and Complex Genome Analysis* Singapore: World Scientific (in press).

Waterston, R. et al. (1992). A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genetics* 1: 114-123.

White, O., C. Soderlund, P. Shanmugam, and C. Fields (1992). Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. *Plant Molecular Biology* 19: 1057-1064.

White, O. and T. Dunning (1993). Computational tools for DNA sequence analysis. In: M. Adams, C. Fields, and J. C. Venter (Eds), *Automated DNA Sequencing and Analysis*. London: Academic Press. pp. 299 - 306.

White, O., T. Dunning, G. Sutton, M. Adams, J. C. Venter, and C. Fields (1993). A quality control algorithm for DNA sequencing projects. *Nucl. Acids Res.* 21: 3829-3838.

Reprints / preprints — removed

DATE
FILMED
10/17/94

END