Data Mining and the Human Genome

Contributors Include:
Henry Abarbanel
Curtis Callan
William Dally
Freeman Dyson
Terence Hwa
Steven Koonin
Herbert Levine
Oscar Rothaus
Roy Schwitters
Christopher Stubbs
Peter Weinberger

January 2000

JSR-99-310

Approved for public release; distribution unlimited.

JASON
The MITRE Corporation
1820 Dolley Madison Boulevard
McLean, Virginia 22102-3481
(703) 883-6997

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information estimated to average 1 hour per response, including the time for review instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 7, 2000	

4. TITLE AND SUBTITLE

Data Mining and the Human Genome

5. FUNDING NUMBERS

13-958534-04

6. AUTHOR(S)

H. Abarbanel, C. Callan, W. Dally, F. Dyson, T. Hwa, S. Koonin, H. Levine, O. Rothaus, R. Schwitters, C. Stubbs, P. Weinberger

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The MITRE Corporation
JASON Program Office
1820 Dolley Madison Blvd
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER

JSR-99-310

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Department of Energy
Biological and Environmental Research
1901 Germantown Road
Germantown, MD 20874-1290

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

JSR-99-310

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

Distribution Statement A

13. ABSTRACT *(Maximum 200 words)*

As genomics research moves from an era of data acquisition to one of both acquisition and interpretation, new methods are required for organizing and prioritizing the data. These methods would allow an initial level of data analysis to be carried out before committing resources to a particular genetic locus. This JASON study sought to delineate the main problems that must be faced in bioinformatics and to identify information technologies that can help to overcome those problems. While the current influx of data greatly exceeds what biologists have experienced in the past, other scientific disciplines and the commercial sector have been handling much larger datasets for many years. Powerful datamining techniques have been developed in other fields that, with appropriate modification, could be applied to the biological sciences.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	None	SAR

Contents

1 EXECUTIVE SUMMARY

As genomics research moves from an era of data acquisition to one of both acquisition and interpretation, new methods are required for organizing and prioritizing the data. These methods would allow an initial level of data analysis to be carried out before committing resources to a particular genetic locus. This JASON study sought to delineate the main problems that must be faced in bioinformatics and to identify information technologies that can help to overcome those problems. While the current influx of data greatly exceeds what biologists have experienced in the past, other scientific disciplines and the commercial sector have been handling much larger datasets for many years. Powerful datamining techniques have been developed in other fields that, with appropriate modification, could be applied to the biological sciences.

Clearly there is a need for more bioinformaticists, as well as computer scientists and engineers who are willing to become involved in bioinformatics research. An ample talent pool already exists from which to recruit those individuals. The DOE can facilitate cross-fertilization between biologists and the non-biological datamining community by sponsoring joint workshops, offering research fellowships to computer scientists who are interested in biological applications, providing access to the unclassified resources of the Advanced Strategic Computing Initiative, and taking advantage of the commercial sector's willingness to make datamining tools freely available to the academic community.

Greater emphasis must be placed on closing the loop between algorithmic analysis and experimental validation. This will require close cooperation between computer scientists and biologists. The DOE should support the development of experimental methods for validating bioinformatics algorithms and the establishment of statistical tests that can be used to assess the robustness of those algorithms. The DOE should take responsibility for

ensuring the provenance of the primary data from the major sequencing centers and making that data freely available in a generic database format with minimal annotation.

2 BACKGROUND

The Human Genome Project, and genomics research in general, is moving at a rapidly accelerating pace. As recently as two years ago there was some doubt as to whether the central goals of the Project would be realized. Those goals, as first articulated in April 1990 [1], were to obtain the complete sequence of the human genome by the end of 2005, and to do so with an accuracy of >99.99% per nucleotide position. This would allow the identification of all $\sim$ 100,000 genes in the human genome and the recognition of the chromosomal location of each of those genes. The total cost of the project (U.S. portion) was estimated to be about $2.5 billion and the final product would be $\sim$ 3 billion base pairs of "finished" human DNA sequence.

In May 1998, the genomics community was rocked by the announcement of a new private venture, born out of a collaboration between The Institute for Genomic Research and Perkin-Elmer Corporation, with the stated aim of sequencing the entire human genome by 2001. A new company named Celera Genomics was founded to carry out this effort. Celera's approach [2] relies on a whole-genome shotgun sequencing strategy, as has proven effective for the sequencing of microbial genomes. Whether this strategy can be scaled up by three orders of magnitude to the size of the human genome and whether it can generate a complete sequence is a matter of debate. Celera plans to take advantage of recent improvements in DNA sequencing technology involving capillary gel electrophoresis sequencing instruments and associated improvements in gel loading and gel reading methods. Specifically, they will utilize 230 of the new ABI PRISM 3700 automated DNA sequencers, each with a potential throughput of 480 kilobases (kb) of raw sequence data per day. Operating 230 instruments at this pace for one year would generate more than 10-fold sequence coverage of the human genome.

In July 1998, JASON conducted a DOE-sponsored study pertaining to functional genomics and research opportunities for the period that will follow acquisition of the human genome sequence (JASON Report JSR-98-315). That study also included an examination of ongoing developments

in genome sequencing technology and the impact of Celera's efforts on the publicly-funded Human Genome Project. The 1998 Report made the following recommendations with regard to genome sequencing:

1. Capitalize on and complement Celera's efforts by addressing the shortcomings of the total shotgun sequencing approach. That approach, even if successful, will likely leave a substantial number of sequence gaps, most of which will be worth closing.

2. Adopt a shotgun sequencing strategy at the level of individual BAC (bacterial artificial chromosome) clones. Each BAC clone contains $\sim$ 150 kb of inserted DNA, providing a manageable assembly problem while allowing greater sequencing throughput compared to the traditional directed sequencing approach.

3. Transition to capillary gel electrophoresis sequencers as soon as possible. These high-throughput instruments will be beneficial for the BAC shotgun sequencing strategy and will help meet the seemingly insatiable demand for increased sequencing capacity.

4. Continue advanced technology development, including sequencing methods that are not based on gel electrophoresis. More so than any other agency, the DOE has the capability to foster long-term technology development in the area of DNA sequencing.

In October 1998 the NIH and DOE issued revised goals for the Human Genome Project [3]. As an interim goal, a "working draft" of the genome would be completed by the end of 2001, aiming for >90% sequence coverage and an accuracy of >99% per nucleotide position. This draft sequence would contain numerous sequence gaps but few physical gaps, and would serve as a platform to anchor the finished sequence. The ultimate goal would still be to obtain the complete sequence of the human genome with an accuracy of >99.99% per nucleotide position. However, the deadline for reaching that goal was pushed forward to the end of 2003. This accelerated timetable necessitated a more focused approach, relying on five major centers to carry

out the bulk of the sequencing effort: the DOE Joint Genome Institute, NIH-sponsored centers at Baylor College of Medicine, The Whitehead Institute, and Washington University School of Medicine, and the Sanger Centre in the U.K. Each center would be free to pursue whatever sequencing strategy it found most productive.

For its part, the DOE Joint Genome Institute is moving aggressively toward adopting new technologies that provide increased sequencing throughput. The three centers that make up the Institute (Lawrence Berkeley, Lawrence Livermore, and Los Alamos National Laboratories) are moving towards a BAC-oriented shotgun sequencing strategy. The Production Sequencing Facility in Walnut Creek, CA is converting entirely to capillary gel electrophoresis sequencers. Based on the results of a side-by-side comparison, the Institute has chosen to purchase the Molecular Dynamics MegaBACE 1000 instrument rather than the ABI PRISM 3700.

In March 1999, the NIH and DOE announced a further accelerated timetable for producing the working draft of the human genome, aiming for completion in Spring of 2000. In July 1999 the NIH announced that three additional centers would be joining the final sequencing campaign: Genome Therapeutics Corporation, Stanford University, and the University of Washington. It is becoming clear that, regardless of what Celera accomplishes in the next few years, the publicly-funded effort will generate a vast amount of human genome sequence data and is likely to complete the sequencing of the human genome well before the original Project deadline.

With most attention having been focused on the monumental task of the sequencing effort itself, there is now a growing concern in the genomics community over problems of data storage and data analysis. The favorite metaphor is that of a tidal wave, with biologists drowning in the onrush of data [4]. These data are derived not only from genome sequencing, but also from sequence annotation, sequence comparison, polymorphism analysis, gene expression analysis, and structure-function studies. As the field moves from an era of data acquisition to one of both acquisition and interpretation, new methods are required for organizing and prioritizing the primary

data so that an initial level of data analysis can be carried out before committing resources to a particular genetic locus. The present JASON study sought to delineate the problems that must be faced in bioinformatics and to identify information technologies, from either within or outside the genomics community, that would be useful in helping to overcome those problems.

3 ACTIVITIES IN BIOINFORMATICS

Bioinformatics is loosely defined as the science of database management and data analysis pertaining to various types of biological information. In the area of genomics, the main activities in bioinformatics are the following:

1. Sequence assembly – determination of a continuous path across many individual DNA sequence reads ($\sim$ 0.5 kb each) and resolving any ambiguities in the data.

2. Sequence annotation – deposition of the assembled sequence into a database, accompanied by information pertaining to the source, quality, and content of the data.

3. Gene finding – analysis of DNA sequence data for indications of an open reading frame that may correspond to an expressed protein.

4. Analysis of non-coding regions – recognition of regulatory elements, inserted sequence elements, and structural features of the chromosome.

5. Sequence comparison – pairwise alignment of either DNA or protein sequences and determination of the degree of similarity between those aligned sequences.

6. Polymorphism analysis – statistical analysis of sequence variation among individuals in a population and correlation of that variation with differences in phenotype.

7. Phylogenetic analysis – sequence comparison between organisms at the level of genes, gene families, or genomes, with the aim of understanding evolutionary relationships.

8. Gene expression analysis – measurement of mRNA or protein expression levels correlated to differences in cell state or environmental conditions.

9. Prediction of RNA and protein structure – computational analysis of primary sequence data leading to recognition of secondary and tertiary structural motifs.

10. Prediction of protein function and interactions – computational analysis of sequence and structural data to infer the functional properties of the corresponding protein.

In conducting this study, JASON heard from investigators working in each of the ten areas listed above. The study did not seek to undertake a review of scientific progress in these areas. Rather, the aim was to determine what computational tools the investigators felt were needed in their research and to consider how those tools might be developed. This Report will focus on four areas chosen from the above list of bioinformatics topics: gene finding, sequence comparison, phylogenetic analysis, and gene expression analysis. These will suffice to illustrate the main conclusions of the study.

3.1 Gene Finding

There are two general approaches to gene finding that currently are in widespread use. One is exemplified by the program GRAIL (Gene Recognition and Assembly Internet Link), which employs a neural network model that has been trained on known genes and can be used to "predict" new genes based on inherent regularities in their primary sequence [5]. GRAIL has undergone several revisions that have added progressively more context-dependent information. The most recent incarnation, termed GRAIL-EXP, includes pattern matching to the database of ESTs (expressed sequenced tags). The other main approach to gene finding employs a hidden Markov model, as exemplified by the program Genie [6]. The model defines a number of discrete states (that in fact may not be "hidden"), corresponding to DNA sequence elements, such as triplet codons or intron-exon junctions. An input DNA sequence is used to drive a succession of state transitions, with the probability of a particular transition being established by training data that reflects the inherent regularity of known genes.

Both neural network and hidden Markov models have been successful in identifying novel genes. They are widely used by biologists, who tend to have little knowledge of the underlying algorithms. These procedures do have some shortcomings. First, they are only weakly informed by the biology. The nodes in the neural network and the states in the Markov model have superficial relevance to actual biological parameters and few attempts have been made to overlay other sources of biological information on these models. A notable exception is GRAIL-EXP, which steps out of the "black box" by including validated experimental data in the model. A second shortcoming is that these models tend to ignore long-range DNA sequence information, which may be critical in recognizing valid open reading frames. The inputs to the algorithm typically consist of consecutive nucleotides or short oligonucleotide strings, with long-range sequence information being treated as an implicit property of the model. A third shortcoming is that the models ignore post-transcriptional changes, such as alternative splicing, RNA editing, ribosomal frameshifting, and post-translational modification, that can significantly alter the meaning of the primary sequence.

Other shortcomings of neural network and hidden Markov models pertain to the algorithms themselves. In some cases the training procedure does not achieve convergence. These cases tend to be rejected from further consideration, but may be informative in understanding the limitations of the model. Finally, insufficient attention is given to the statistical significance of the results that are obtained. Few statisticians are involved in bioinformatics research, even though they could make a significant contribution.

While the developer of a gene finding method usually is aware of its shortcomings, the end users tend to utilize the method uncritically and in a turn-key fashion. In further improving gene finding methods it will be important to close the loop between bioinformaticists and experimentalists. When a predicted gene is either validated or falsified experimentally, that result should be communicated back to the developer who can use it to improve the search algorithm. Ideally the feedback would occur in a semi-automated manner, for example, by including a self-reporting feature in a web-based bioinformatics algorithms. It would be beneficial to develop reliable meth-

ods for experimental validation of predicted genes, especially those that could be implemented on a large scale at major sequencing centers. For example, a family of expression cassettes could be developed that are compatible with the sequencing vectors and that allow for a rapid test of gene expression in various standard cell lines.

As the number of experimentally validated genes increases, this validated information will become more useful in augmenting the gene finding procedure. More so than the EST database, the database of validated full-length genes will aid in assigning confidence values to predicted genes. Other experimental data pertaining to alternative splicing and other forms of post-transcriptional modification would be useful as well when accessed through a look-up table that complements the primary search algorithm. Closing the loop between software specialists and experimentalists will reduce the barrier between their respective scientific disciplines, causing both groups to focus on the common goal of making gene finding methods more inclusive of the relevant biology and therefore more accurate.

3.2 Sequence Comparison

The preferred algorithm for DNA and protein sequence comparison is BLAST (Basic Local Alignment Search Tool). This method was first developed in 1990 [7] and has undergone several refinements since then. A BLAST search involves submission of a query sequence that is compared against all entries in the available sequence databases. The query sequence is aligned against each test sequence, in most cases allowing for the possibility of insertions or deletions. Positive matches are reported and annotated with the score for the number of identical and similar amino acid residues as well as the probability that such a match would occur by chance. The most recent versions of BLAST allow for either an iterated search that takes into account the best matches found in the previous round (PSI-BLAST) or a focused search that places special value on sequence patterns that occur within the query sequence (PHI BLAST).

BLAST is likely to be the most widely utilized piece of bioinformatics software. It is used to draw inferences regarding the function of newly-identified genes and to search for homologous genes within a gene family or among different organisms. The results of a typical BLAST search tend to include many false positives, caused by the intrinsic sequence regularity of proteins. Most users gladly accept a high rate of false positives because it gives them more "hits" to consider and a lower chance of missing the genes of marginal sequence similarity that might have high biological significance. However, as the database expands and the number of query sequences increases, there will be less tolerance for false positives. More attention will need to be directed toward understanding the sensitivity function associated with BLAST and other sequence comparison procedures. It would be useful, for example, to randomly degrade the search string in a way that reflects normal sequence statistics, and determine which of the matched sequences are the first to drop out. The presumption, which remains to be tested, is that true matches will tend to outlast false positives as the information is progressively degraded. It also would be useful to investigate procedures for normalizing the matching function with respect to statistical regularities in protein sequence and structure.

Perhaps the greatest shortcoming of BLAST and related methods is that they have weak predictive value for regulatory regions and other non-coding portions of the genome. This is a difficult problem because non-coding regions may be important for understanding function, but tend not to have enough sequence regularity to allow meaningful sequence alignments. One recourse would be to incorporate phylogenetic data, utilizing sequence information from several different organisms and normalizing for their phylogenetic distance. At present this is difficult to do because of the paucity of completed genome sequences. However, the next decade will see an explosion in the accumulated sequence data, not just from expressed genes but from intergenic regions, which will facilitate sequence comparisons involving non-coding regions.

3.3 Phylogenetic Analysis

In the near future there will be a large number of fully sequenced genomes, covering a diverse range of species representing all major branches of the tree of life. Techniques for sequence comparison will be applied not only to a single query sequence in relation to the sequence database, but also to all of the pairwise comparisons that can be made among a set of query sequences. These global comparisons will be used to infer phylogenetic relationships and reconstruct evolutionary history. Phylogenetic analysis therefore involves combinatorial pairwise alignment, a task that has a computational cost of the order $O(n^k)$, where n is the length of the sequence and k is the number of sequences being compared.

More so than most other areas of bioinformatics, phylogenetic analysis can benefit from access to high-performance, multi-processor computing. In the 1998 JASON Report it was suggested that the DOE should assume a strong role in comparative genome sequencing. The DOE already has established its place in comparative genomics by supporting the sequencing of archael and eubacterial genomes, and comparative genomics is more closely associated with studies of biological diversity than biomedical applications. The DOE has in place substantial computing resources within the context of the Accelerated Strategic Computing Initiative (ASCI). It is natural to suggest that these resources might be leveraged to assist in the development and application of algorithms for comparative sequence analysis.

The three high-performance computing platforms, ASCI Red at Sandia, Blue Pacific at Lawrence Livermore, and Blue Mountain at Los Alamos National Laboratories, have projected capabilities in the range of 1–3 teraOps. The ongoing commitment to ASCI, in support of the stockpile stewardship program and other defense-related needs, ensures that the existing platforms will be replaced over the next few years by even more powerful computers. This will create an opportunity for making the previous-generation machines available for unclassified use. The DOE has established a model for providing open access to high-performance computing through its operation of the

ACL Nirvana Machine at Los Alamos National Laboratory. Bioinformaticists who are carrying out phylogenetic analysis, structure prediction, and other computationally intensive tasks have benefited and should continue to benefit from access to high-performance computing at the ASCI facilities.

3.4 Gene Expression Analysis

One of the most rapidly progressing areas of research in bioinformatics concerns the analysis of gene expression. Powerful experimental systems have been devised that allow one to measure the expression level of all known genes in an organism under a defined set of cellular conditions. This is accomplished using expression arrays that contain spatially-ordered DNA probes corresponding to each of the genes being interrogated. Total messenger RNA is harvested from the cells, labeled during reverse transcription, and allowed to hybridize to complementary probes on the array. The intensity of the label at each probe position, typically normalized to a control sample, reflects the expression level of the corresponding gene. The power of expression arrays lies in their high degree of parallelism. A single "gene chip" that is only 1.8 cm^2 may contain 30,000 DNA probes that can be addressed simultaneously. The expression state of the cell can be monitored over time in relation to cellular events, such as the progression of the cell cycle or the application of an external stimulus.

Expression analysis has the potential to generate vast amounts of data. Computational techniques have been developed to organize this data and mine it for inherent regularities. One such technique is cluster analysis [8], which has been employed in many other areas of data analysis. Clustering algorithms may be either referenced to an established hierarchy or allowed to self-organize according to some statistical model. When applied to gene expression data, the clustering algorithm typically is unsupervised and is based on simple pairwise comparisons that result in hierarchical clusters. The organization of the clusters is not pre-specified, but there is nonetheless a strong observed tendency for genes of related function to fall within the

same cluster. This behavior can be valuable in inferring the function of an unknown gene based on its propensity to cluster with other genes of known function.

One of the shortcomings of cluster analysis as currently applied to gene expression data is that there is no established metric for the robustness of a cluster. The penalty for expanding or collapsing a cluster is not known, nor is the cost function that would allow a gene to belong to more than one cluster simultaneously. This shortcoming could be addressed either computationally or experimentally by seeking to determine what makes a cluster break. The computational approach would involve randomly degrading the dataset and observing how the clustering pattern is affected as correlations in the expression pattern becomes more tenuous. Such an exercise assumes, perhaps incorrectly, that valid clusters will be more resistant to degradation of the dataset compared to invalid clusters. The experimental approach would involve reducing the expression of one or more genes in a cluster (e.g. by antisense inhibition) and determining how the structure of that cluster and other clusters in the hierarchy are affected. Such exercises will help to define the limits of what constitutes a valid cluster. They may also lead to the realization that multi-parent clustering algorithms are needed to provide a more accurate description of how gene expression is organized.

As with gene finding methods, expression analysis could be improved by striving to close the loop between computation and experimentation. Predictions made by clustering algorithms should be validated experimentally and the validated examples should be used to refine the clustering algorithms. A high-throughput method, such as a two-hybrid system, could be used to confirm that some genes in a cluster are indeed functionally related. Information about the chromosomal location of genes might also assist in the interpretation of expression data. Expression patterns could be "spiked" by the inclusion of genes that are expressed on a plasmid. If the clustering algorithms are behaving properly, two plasmid genes that are expressed from a common promoter will cluster together very tightly. In general, cluster prediction should be treated as an iterative process that requires the close interaction of bioinformaticists and experimentalists.

4 DATA MINING

Biologists are experiencing a level of data influx that greatly exceeds what they have seen in the past. This is true in many areas of biology, but especially in the genome sciences. There is much hand-wringing at present and a growing sense that something must be done to change the culture of biology so that it can cope with the tremendously increased volume of data. The private sector, including most major pharmaceutical companies and many biotechnology companies, has made a significant investment in bioinformatics. Several major universities have initiated research and training programs in this area and more are on their way. Individuals calling themselves "bioinformaticists" are in great demand today and will continue to be so in the years ahead.

The NIH Working Group on Biomedical Computing recently issued an Advisory Committee Report to the NIH Director on the subject of biomedical information science and technology [9]. That Report made the following specific recommendations:

1. Establish 5–20 National Programs of Excellence in Biomedical Computing to help educate researchers in this area.
2. Establish a new NIH-sponsored Program on information storage, curation, analysis, and retrieval (ISCAR).
3. Increase funding for basic research in biomedical computing through R01 grants.
4. Foster a scalable national computer infrastructure.

These recommendations are sensible and appropriately place strong emphasis on the need to train individuals with skills in both the computational and biomedical sciences. The context of the Working Group Report extends beyond bioinformatics pertaining to genome sciences and includes all areas of biomedical computing. However, it is genomics issues that are driving

much of the discussion and the impending availability of the complete human genome sequence that is creating much of the sense of urgency.

4.1 Sociological Issues

While these are data-intensive times for biologists, it is important to note that other scientific disciplines have been handling much larger datasets for many years. Astronomers, climate modelers, hydrodynamicists, and structural engineers, for example, have been dealing with terabytes of information, while bioinformaticists must face "only" gigabytes of DNA sequence data and gene expression data. Outside the sciences the amount of information being gathered and interpreted is often even greater. In the commercial sector, for example, data from billions of credit card transactions and cash register receipts are processed annually in order to analyze purchasing patterns and uncover evidence of fraud. In the national security arena, vast amounts of electronic and image data are processed in close to real time and mined for features that are of special interest.

The above discussion is not meant to diminish the significance of problems that must be faced in bioinformatics, but to point out that these problems are likely to be manageable. There are no "hard" computational tasks in bioinformatics that would exceed the capabilities of modern computers. Adequate processing power and data storage capacity are available to meet these needs. High-speed internet connections and high-density physical media allow easy transfer of large amounts of data. Heightened concern over bioinformatics issues is appropriate, but the present era should be viewed more as one of opportunity than of serious obstacles.

If managing and analyzing vast amounts of data has become routine in so many other areas, why haven't the relevant techniques crossed over to the biological sciences? There are several reasons for this. First, the interpretation of biological data requires special knowledge of biochemistry, molecular biology, and cell biology. (A similar statement could be made for most other fields.) Second, most biologists interact with computers in a

"point-and-click" fashion and have little training in the information sciences. Third, few individuals exist who have expertise in both computer sciences and biology and, therefore, are in a position to foster closer interaction between the two disciplines. Fourth, the rush to obtain the complete human genome sequence and other primary sequence data has left little time for analysis. Fifth, the techniques for data acquisition are changing rapidly, making it uncertain which bioinformatics methods will have lasting value.

Clearly there is a need for more bioinformaticists. Individuals who are just as comfortable thinking about hidden Markov models as transcriptional activation will have a profound impact on biology over the next decade. In addition, there is a significant need for database engineers, computer scientists, statisticians, and applied mathematicians who are willing to become involved in bioinformatics. An ample talent pool from which to draw those individuals already exists, for example, in the so-called KDD (Knowledge Discovery in Databases) community. Persons working in this area usually have no formal training in biology and would not be capable of initiating a Program of Excellence in Biomedical Computing as envisioned by the NIH Working Group. But many of these individuals would rather apply their craft to human genome sequence data than, say, cash register receipts. They should be recruited to participate in bioinformatics research, as will be discussed below.

The KDD community is large and intellectually rigorous. It has a peer reviewed journal, *Data Mining and Knowledge Discovery*, and holds regular meetings ranging from small workshops to an annual international conference. Of the 52 papers that were chosen for presentation at the upcoming "Fifth International Conference on Knowledge Discovery and Datamining", none pertain to bioinformatics. This would change if members of the KDD community were allowed to integrate with genome scientists as part of a bioinformatics research program. Like the physicist who assists a group of NMR spectroscopists or the veterinarian who participates in pharmacology research, professional dataminers have much to contribute if allowed to work as equal partners with biologists who are engaged in large-scale data analysis.

The DOE can facilitate cross-fertilization between biologists and dataminers in several ways. First, it could sponsor a workshop to bring together members of the two communities, not to design algorithms, but to talk about what is needed in biology and what the various datamining approaches have to offer. Second, the DOE could offer research fellowships to computer scientists who have an interest in biological applications, even if they are not well versed in biology. These individuals would be expected to work as part of a research team that is handling large amounts of biological information. Third, ASCI resources that are available for unclassified use could be offered as an inducement to attract dataminers to the sequencing centers and other DOE-supported laboratories. Fourth, the DOE should take advantage of the commercial sector's willingness to make datamining tools available to the academic community. For example, the IBM *Intelligent Miner* software package can be obtained free of charge for use in scientific research. This would also help to build ties to the commercial world of datamining.

Members of the datamining community who have the courage and desire to enter bioinformatics will require some special considerations. They need to be supported in their efforts to learn the concepts of biology and techniques of genomics research. They should not be segregated in dark computer-filled rooms, but made to enter the mix with other members of the research team. They should not be expected to provide computer support services for the laboratory; this activity should be viewed as part of facilities management and not bioinformatics. Finally, and perhaps most sensitively, they require a salary that is comparable to what other computer scientists receive. In the current job market, this will be higher than the salary of molecular biologists who have the same number of years of experience.

4.2 Database Organization

An important area of concern is the current lack of standardization in genome databases. The *de facto* standard at present is GenBank, which is an annotated collection of all publicly available DNA sequence data, currently

consisting of more than 3.4 billion bases in more than 4.6 million sequence records. In reality, GenBank is little more than a community data dump. There is tremendous variability in the accuracy, completeness, and degree of annotation of GenBank entries. While it is useful to have all of the sequence information gathered in one location, the uneven quality of the data reduces its usefulness for many bioinformatics purposes.

Prior DOE efforts to build a definitive genome database have failed for both scientific and sociological reasons. It was premature then, and still is premature, to try to impose a database standard on a field that is changing so rapidly. Ultimately the scientific community will determine its own standards and specifications. However, there are two important steps that the DOE can take now to prepare for that eventuality. First, it can help to guarantee the provenance of the primary data, especially from the major sequencing centers. This includes the raw sequence data and the assembled sequences that are part of both the draft and finished sequencing efforts. Second, the DOE can make that primary data freely available to the scientific community in the form of a generic database with minimal annotation. This information should conform to a low-level format standard, encompassing the raw data, statistics regarding its accuracy, and annotation pertaining to the source of the data and its context within the genome.

As discussed in the 1998 JASON Report, a modular approach is needed for database management so that the data gathering functions performed by experimentalists are separated from the cataloguing and data manipulation functions performed by data analysts. This will enable one group of investigators to focus on data acquisition in the face of changing research methods while the other group focuses on data management in the face of changing computer technology. So long as the integrity of the primary data is assured, it will be possible to translate that data into different formats at a later date. By warehousing and distributing the primary data, the DOE can facilitate a modular approach to database management and allow a broader range of investigators to participate in the analysis of genomics data. It is important to avoid the attitude that data generated at one of the sequencing centers somehow belongs to that center. The centers are funded to generate data for

the scientific community and are obliged to support a policy that allows the broadest possible distribution of that data.

As an increasing number of individuals try their hand at developing bioinformatics tools, it will be important to have an objective means of assessing the utility of those tools. End users should be encouraged to utilize tried-and-true methods whenever possible. The DOE should support the development of benchmark datasets that can be used to evaluate new algorithms. If a new algorithm is to replace an existing one, then it must first demonstrate its superiority when applied to the benchmark data. The process of certifying new hardware or software designs through the use of performance benchmarks is well established in computer engineering. The developers of protein structure prediction methods have taken a similar approach, regularly holding contests to compare the performance of the various algorithms. A situation to avoid is one in which each sequencing center analyzes its own experimental data according to its own favorite algorithm. The preferred situation is one in which all centers make their primary data freely available and many different groups compete to develop algorithms that best analyze the data. Methods that prove most efficacious can then be offered for general use.

5 RECOMMENDATIONS

This is an exciting time in the biological sciences. The acquisition of the complete sequence of the human genome is close at hand. Additional data are pouring in pertaining to the annotation, analysis, and interpretation of that sequence information. Academic, government, and commercial organizations are scrambling to develop and implement methods to process the data. Now that the primary goals of the Human Genome Project appear to be within reach, concern has shifted to the question of how the research community will handle the data. Appropriately, the cry has gone out to train more computational biologists who can develop software tools to meet the needs of genomics research.

Without diminishing the significance of the current shortfall of bioinformaticists, two assuasive statements should be made. First, the sky is not falling. The problems that must be faced in genomics are manageable and similar to those that have been addressed in other areas of data analysis. Second, do not expect too much from computers. Bioinformatics methods produce leads, not answers. It will be necessary to validate those leads experimentally and, whenever possible, use the validated information to refine the predictive algorithms.

This study did not seek to review scientific progress in the various areas of bioinformatics, but to consider technical and sociological issues that apply more broadly to problems of data handling and data analysis in genomics research. In general, the picture is a positive one, dominated by a sense of great opportunity. The following summary recommendations are meant to suggest how the DOE can play an active role in helping to realize that opportunity.

1. Support the development of high-throughput methods for experimental validation and subsequent refinement of bioinformatics algorithms.

2. Support the development of statistical tests to assess the robustness of bioinformatics algorithms.

3. Allow bioinformaticists access to the non-classified resources of the Advanced Strategic Computing Initiative.

4. Increase training opportunities for individuals skilled in both computer science and experimental biology.

5. Recruit members of the datamining community by supporting joint workshops, fellowships, and the greater use of commercial datamining tools.

6. Do not try to impose a standard for database organization and management.

7. Ensure the provenance of the primary data from the major sequencing centers and make that data freely available in a generic database format with minimal annotation.

8. Support the development of benchmark datasets to assist in the evaluation of new bioinformatics algorithms.

It is important to close the loop between algorithmic analysis and experimental validation. This will require close cooperation between computer scientists and biologists. Training opportunities should be provided for individuals seeking joint skills in these areas. In the meantime, the substantial number of people working in the non-biological datamining community provides a talent pool from which to draw upon. Despite their lack of expertise in the biological sciences, these individuals should be directed to work side-by-side with experimentalists who are engaged in the validation and implementation of bioinformatics tools. The data analysts need not be affiliated with the data gatherers. A modular approach that separates data gathering from analysis and a distribution system that makes the primary data freely available to the scientific community will allow many laboratories to participate in data analysis.

References

[1] U. S. Department of Health and Human Services and Department of Energy, *Understanding Our Genetic Inheritance. The U. S. Human Genome Project: The First Five Years*, 1990.

[2] Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. & Hunkapiller, M. *Science* **280**, 1540–1542, 1998.

[3] Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L. & members of the DOE and NIH planning groups *Science* **282**, 682–689, 1998.

[4] Reichhardt, T. *Nature* **399**, 517–520, 1999.

[5] Uberbacher, E. C. & Mural, R. J. *Proc. Natl. Acad. Sci. USA* **88**, 11261–11265, 1991.

[6] Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. *Proceedings of the Conference on Intelligent Systems in Molecular Biology '96* (States, D. J., Agarwal, P., Gaasterland, T., Hunter L. & Smith, R., Eds.) AAAI/MIT Press, pp. 134–142, 1996.

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. *J. Mol. Biol.* **215**, 403–410, 1990.

[8] Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868, 1998.

[9] Botstein, D., Smarr, L. *et al. The Biomedical Information Science and Technology Initiative*, Working Group on Biomedical Computing Advisory Committee to the Director, National Institutes of Health, 1999 (available at www.nih.gov/welcome/director/060399.htm).

A APPENDIX – Algorithmic Methods in Bioinformatics

The enormous success of molecular biology has resulted in the production of vast amounts of data, leading biologists to seek automated ways of parsing and interpreting that data. DNA sequence data is time independent and typically is interpreted by establishing a mapping from the primary sequence to the genes that it encodes. Gene expression data typically are time dependent and are interpreted as a set of signals that change over time.

In the case of DNA sequence data, one starts with an ordered set of symbols (e.g. the four nucleotides or the 64 triplet codons) drawn from the list:

$$X = x_1, x_2, \ldots, x_K,$$

and attempts to find the corresponding set of genes:

$$G = g_1, g_2, \ldots, g_L.$$

In other words, one wishes to find the mapping:

$$G = M(X),$$

from the K-dimensional space of sequence elements to the L-dimensional space of associated genes. This situation is fundamentally the same as that encountered in database-model associations in essentially all fields of science. Nothing is special about bioinformatics except that the data is to be interpreted in terms of genetic sequences and functional macromolecules.

If the data are time dependent, then one is seeking a time evolution relating a signal $x_k(t)$ at time t to the signal at some later time. There are typically many signals k, so one is interested in determining the dynamical mapping:

$$x_k(t+1) = F_k((x_1(t), x_2(t), \ldots x_K(t), \mu); k = 1, 2, \ldots K,$$

where μ is a set of parameters defining the unknown function $F_k(\bullet, \mu)$.

Finding the map $M(\bullet)$ from observed data on X and G or the function $F(\bullet)$ from observed data on $X(t)$ is a standard problem. In bioinformatics two approaches to this problem are most commonly taken. The first uses a universal class of approximation functions termed "neural networks" that express essentially any smooth function such as $X = M(G)$ in terms of a basis set of "sigmoidal" functions:

$$x_k = \sum_{m=1}^{M} w_m \, \sigma_m \, (\sum_{n=1}^{N} T_{kn} \, g_n + b_k).$$

The functions $\sigma(\bullet)$ are typically logistic or atanh, and the parameters $\mu = \{w_n, T_{nj}, b_n\}$ characterize the model. When one selects the number N, J of "neural units" and applies training data on the set of X's and the set of G's, the parameters can be determined by various means and a model that includes these parameters can be developed. The model is then used on a new set of sequence data to "predict" the associated genes.

Because this representation of nonlinear functions is universal, when N and J are large enough, it is clear that one can always use such a representation to achieve better and better predictions for larger and larger sets of parameters and larger and larger sets of validated data on which to train those parameters. What is critically missing from this black box approach is the interpretation of the myriad of parameters in terms of some biological function. Furthermore, the number of parameters can be quite large, from several hundreds to several thousands, and the biological justification for this large number remains slightly mysterious. How should one interpret the use of models such as large neural networks in mapping DNA sequences to genes? Perhaps they can be seen as a guide to the user, suggesting what experiments might be done to establish that certain sequences do in fact represent genes.

Similar neural network models have been applied to time series data, for example, gene expression over the course of development. In this work, difference equations for maps in discrete time or differential equations for developmental changes in continuous time are represented by the neural network and the relevant data parameters are selected by some fitting method. In one example, Eric Mjolsness and colleagues [Mjolsness, E., Mann, T., Castano, R. & Wold, B. *JPL Technical Report JPL-ICTR-99-4*, 1999] have

attempted to determine the "intensity" of expression $Y_j(t)$ of a gene in terms of a time constant for the decay of that intensity and a feedback from the activity of both that gene and other genes. This is expressed as:

$$dY_j(t)/dt = F(\sum_{m=1}^{M} T_{jm}Y_m(t) + b_j),$$

where $F(\bullet)$ is another sigmoidal function, chosen for convenience and familiarity rather than for any association with the biological processes being described. Employing this type of model and determining the function's coefficients with training data, Mjolsness and colleagues were able to "predict" the expression time series of other genes.

Another common approach in bioinformatics involves the use of Hidden Markov Models (HMMs) to derive the association between sequence and function. When applied in this way, the models typically are probabilistic and make the assumption that the present state of the system is independent of prior or future states. For example, states that correspond to individual nucleotides in a DNA sequence are assumed to be independent of previous or subsequent nucleotides. The introduction of probability in what would seem to be a deterministic process is partly a way of expressing the presence of errors or inherent limits in the data. One replaces definite relations, such as $X = M(G)$, with conditional probabilities for a sequence X given a gene sequence $G : P(X|G)$. As noise and other errors disappear, this probability density becomes a delta function on the appropriate spaces and expresses $X = M(G)$ without ambiguity.

In the HMM approach one associates a succession of state transitions among members of the sequence X with the probabilities that at each state a "symbol is emitted" which can be related to an observable. Davis Haussler and colleagues have employed HMM methods to associate nucleotide triplet sequences with observed protein structure [Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. *J. Mol. Biol.* **235**, 1501–1531, 1994]. In this case, one identifies three types of match states that are positions in three-dimensional space within a protein structure and associates with these match states one of twenty "emitted" amino acids. From a model of the relationship between a sequence of match states and amino acids in a known database,

one tries to establish a means of relating a new protein structure to the corresponding amino acid sequence.

In another application of HMMs, investigators sought to obtain the most likely sequence of genes given an observed sequence of DNA [Kulp, D., Haussler, D., Reese, M. & Eckman, F. *Proceedings of the Conference on Intelligent Systems in Molecular Biology '96* (States, D. J., Agarwal, P., Gaasterland, T., Hunter L. & Smith, R., Eds.) AAAI/MIT Press, pp. 134–142, 1996]. This was done by maximizing the probability $P(X, G|\text{model})$, that probability being the product of the individual probabilities along the sequence of X's. In this way, it was possible to identify 93% of true exons in a dataset. At least 29% of the predicted genes did not match any known genes, although the number of false positives was large.

Studies employing neural networks or HMMs focus on methods for efficiently determining the myriad of parameters in the model. The choice of a particular model is usually made on the basis of the percentages of correct predictions, false positives, and false negatives when applied to novel data sets. While this approach seems rational, it begs the question of what fundamental biological information is gained by determining the model parameters. The authors of one study pointed to this issue directly: "We believe that the answer to the problem of small training sets is to add more prior knowledge into the training process. One way to do this is to start with a better initial model" [Krogh *et al., vide supra*].

The critical issue with neural networks, HMMs, and other black box models is that these methods typically lack a biological rationale for the forms of the models, including the nature of the parameterization used to generate those forms. One hopes to gain some clue as to the nature of the underlying processes, but the connection is usually missing between the vast number of parameters and the underlying biological processes that give rise to those parameters. Also lacking is the suggestion of how biological experiments could be used to verify the main aspects of the model.

It is reasonable to suggest a series of experiments and associated modeling efforts that, roughly, run as follows: (1) choose a set of known gene

sequences and their associated upstream and downstream elements; (2) employing a portion of those sequences, determine the coefficients in a model that incorporates both the kinematic constraints suggested by the model and as much of the relevant biological dynamics as can be built into the model so that the parameters have a clear association with biological observables; (3) verify the model on the remaining portion of the data; (4) apply the model to identify the genes within another set of DNA sequences from the same organism; (5) verify the predictions of the model using standard laboratory methods. While this program sounds obvious, there do not appear to be many examples of it in the literature. It would provide the kind of verification that is necessary if progress is to be made in understanding the biological dynamics that underlie the data. Otherwise one is left with black boxes having many internal parameters that remain black even after the work has been completed.

B APPENDIX – A Cautionary Note Regarding Hidden Markov Models

The use of Hidden Markov Models has become widespread in bioinformatics, although their application in this area is not quite traditional. They appear to be utilized primarily for their computational advantages in deciding which of a large number of parsings of a long string of nucleic acids (or many moderate-length strings) is the correct one based on the fact that it parses with highest probability.

On a hidden or partly hidden Markov model assumption, the probability of a given parsing is readily computed, quite independently of any secondary considerations such as the underlying biology. The answer is given by the optimal point the HMM reaches, using either a standard one-way (Viterbi) or two-way probability maximizing procedure. It is important to note that the optimum found by this procedure is very often a false maximum, hence possibly misleading. The presence of symmetries or functional homologies in the underlying data almost guarantees the existence of false optima. This statement does not refer to local maxima that are not global maxima. It is possible that an HMM gets hung up at such a point, for example, a saddle point. The concern applies to cases in which there are many global maxima, not all of which are biologically relevant.

If one is only looking for the best parsing, the biological significance may be unimportant. But if physical conclusions based on the parameters of the hidden Markov process are to be inferred from the optimum, then false conclusions may be drawn. In order to illustrate this point it is helpful to adopt a notation in which Greek letters represent the hidden states and Latin letters represent the visible states of the process. Let P be the true transition matrix for the Markov process. Thus $P(\alpha, \beta)$ is the conditional probability of β given a particular α. The row vector p is the probability distribution of hidden states. The column vector of all 1's is denoted by δ. Finally the matrix A^i is the diagonal matrix whose α'th diagonal entry is the probability that the hidden state α is read out as the visible state i.

With this notation it is easy to express the probability of seeing the visible string $i_1, i_2, \cdots, i_n$:

$$pA^{i_1}pA^{i_2}\cdots pA^{i_n}\delta.$$

For any non-singular H it is also:

$$pH(H^{-1}A^{i_1}H)(H^{-1}pH)(H^{-1}A^{i_2}H)\cdots(H^{-1}A^{i_n}H)H^{-1}\delta.$$

Whenever H can be chosen so that all of $pH, H^{-1}A^iH$, $H^{-1}PH$, and $H^{-1}\delta$ have the correct form, the result is two HMMs with the same output statistics. ("Correct" in this case means, as appropriate, diagonal, row-stochastic, non–negative, etc.). Notice that any permutation matrix H meets this requirement, but this amounts simply to renaming all of the hidden states.

Suppose, however, that two rows of the transition matrix P are the same, say the first and second rows. From Bayes Theorem, or simply from the definition of conditional probability, it is straightforward to infer that the first two columns also are the same. Thus $H^{-1}PH = P$, where H is the permutation matrix interchanging the first and second states. Then, without altering P, each A^i may be replaced by $H^{-1}A^iH$ to produce an HMM with the same visible statistics.

More generally, if $H^{-1}PH = P$ for some permutation matrix H, then P may be left unchanged and each A^i may be replaced by $H^{-1}A^iH$. Indeed, P may be left unchanged and each A^i may be replaced by $H^{-k}A^iH^k$ for any power k of the original permutation matrix H. Thus the peculiar symmetry $P = H^{-1}PH$ leads to many statistically equivalent HMMs.

It is not clear if nature has tended to build such symmetries into P, the functional units of the gene, assuming that this modeling bears some relation to the true state of affairs. There are some obvious symmetries within the structure of certain genes and regulatory elements. There may be other hidden symmetries, for example, a permutation H such that $H^{-1}A^iH = A^i$ for all i. In this case, the A^i's could be left unchanged and P could be replaced by $H^{-1}PH$.

More substantial perturbations are possible. Let $K = cI + dH$, where I is the identity matrix and $c + d = 1$. Then K commutes with all the A^i's,

so they may be left unchanged and P may be replaced by $K^{-1}PK$, provided only that it and pK have non-negative entries, which is almost surely true for small c or small d. This strongly suggests the possibility of a whole ridge of optima for some HMMs.

The difficulties pointed out here may not be a problem for some datasets in bioinformatics, and in all likelihood can be remedied if one is aware of their possibility. These remarks, therefore, are to be regarded as cautionary.

DISTRIBUTION LIST

Director of Space and SDI Programs
SAF/AQSC
1060 Air Force Pentagon
Washington, DC 20330-1060

CMDR & Program Executive Officer
U S Army/CSSD-ZA
Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0150

Superintendent
Code 1424
Attn Documents Librarian
Naval Postgraduate School
Monterey, CA93943

DTIC [2]
8725 John Jay Kingman Road
Suite 0944
Fort Belvoir,VA 22060-6218

Dr. A. Michael Andrews
Director of Technology
SARD-TT
Room 3E480
Research Development Acquisition
103 Army Pentagon
Washington, DC 20301-0103

Dr. Albert Brandenstein
Chief Scientist
Office of Nat'l Drug Control Policy
Executive Office of the President
Washington, DC 20500

Dr. H. Lee Buchanan, I I I
Assistant Secretary of the Navy
(Research, Development & Acquisition)
3701 North Fairfax Drive
1000 Navy Pentagon
Washington, DC 20350-1000

Dr. Collier
Chief Scientist
U S Army Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0280

D A R P A Library
3701 North Fairfax Drive
Arlington, VA 22209-2308

Dr. Victor Demarines, Jr.
President and Chief Exec Officer
The MITRE Corporation
A210
202 Burlington Road
Bedford,MA01730-1420

Mr. Frank Fernandez
Director
DARPA/DIRO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Mr. Dan Flynn [5]
Deputy Chief
OSWR
CDT/OWTP
4P07, NHB
Washington, DC 20505

Dr. Paris Genalis
Deputy Director
OUSD(A&T)/S&TS/NW
The Pentagon, Room 3D1048
Washington, DC 20301

Dr. Lawrence K. Gershwin
NIO/S&T
2E42, OHB
Washington, DC 20505

General Thomas F. Gioconda [5]
Assistant Secretary for Defense
US Department of Energy
DP-1, Room 4A019
Mailstop 4A-028
1000 Independence Ave, SW
Washington, DC 20585

Mr. David Havlik
Manager
Weapons Program Coordination Office
MS 9006
Sandia National Laboratories
PO Box 969
Livermore, CA94551-0969

DISTRIBUTION LIST

Dr. Helmut Hellwig
Deputy Asst Secretary
(Science, Technology and Engineering)
SAF/AQR
1060 Air Force Pentagon
Washington, DC 20330-1060

Dr. Robert G. Henderson
Director
JASON Program Office
The MITRE Corporation
1820 Dolley Madison Blvd
Mailstop W553
McLean, VA 22102

J A S O N Library [5]
The MITRE Corporation
Mail Stop W002
1820 Dolley Madison Blvd
McLean, VA 22102

Mr. O' Dean P. Judd
Los Alamos National Laboratory
Mailstop F650
Los Alamos, NM 87545

Dr. Bobby R. Junker
Office of Naval Research
Code 111
800 North Quincy Street
Arlington, VA 22217

Dr. Martha Krebs
Director
Energy Research, ER-1, Rm 7B-058
1000 Independence Ave, SW
Washington, DC 20858

Lt Gen, Howard W. Leaf, (Retired)
Director, Test and Evaluation
HQ USAF/TE
1650 Air Force Pentagon
Washington, DC 20330-1650

Dr. Arthur Manfredi
ZETA Associates
10300 Eaton Drive
Suite 500
FairfaxVA 22030-2239

Dr. George Mayer
Scientific Director
Army Research Office
4015 Wilson Blvd
Tower 3, Suite 216
Arlington, VA 22203-2529

Ms. M. Jill Mc Master
Editor
Journal of Intelligence Community Research
and Development (JICRD)
Investment Program Office (IPO)
1041 Electric Avenue
Vienna,VA 20180

Dr. Thomas Meyer
DARPA/DIRO
3701 N. Fairfax Drive
Arlington, VA 22203

Dr. Bill Murphy
ACIS
6014 Martins Landing Lane
BurkeVA 22015

Dr. Julian C. Nall
Institute for Defense Analyses
1801 North Beauregard Street
Alexandria, VA 22311

Dr. Ari Patrinos [5]
Associate Director
Biological and Environmental Research
SC-70
US Department of Energy
19901 Germantown Road
Germantown,MD 207874-1290

Dr. Bruce Pierce
USD(A)D S
The Pentagon, Room 3D136
Washington, DC 20301-3090

Mr. John Rausch [2]
Division Head 06 Department
NAVOPINTCEN
4301 Suitland Road
Washington, DC 20390

DISTRIBUTION LIST

Records Resource
The MITRE Corporation
Mailstop W115
1820 Dolley Madison Blvd
McLean, VA 22102

Dr. Fred E. Saalfeld
Director
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5000

Dr. Dan Schuresko
O/DDS&T
OSA/ATG
Room 23F20N, WF-2
Washington, DC 20505

Dr. John Schuster
Submarine Warfare Division
Submarine, Security & Tech
Head (N875)
2000 Navy Pentagon Room 4D534
Washington, DC 20350-2000

Dr. Michael A. Stroscio
US Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

Dr. George W. Ullrich [3]
ODUSD(S&T)
Director for Weapons Systems
3080 Defense Pentagon
Washington, DC 20301-3080

Dr. David Whelan
Director
DARPA/TTO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Dr. Edward C. Whitman
US Naval Observatory
Nval Oceanographers Office
3450 Massachusetts Ave, NW
Washington, DC 20392-5421

Dr. Peter D. Zimmerman
Science Advisor
ACDA
320 21st Street, NW
Washington, DC 20451