may enhance our ability to understand as well as control quantum systems.

Bob: I thought all the fuss about quantum computing was about engineering—but that sounds like something you'd read in *Science*.

Alice: Nah, they'd never publish something like this.

References

1. L. M. Adleman, *Science* **266**, 1021 (1994).
2. C. H. Bennett, G. Brassard, C. Crepeau, R. Jozsa, A. Peres W. K. Wootters, *Phys. Rev. Lett.* **70**, 1895 (1993).
3. D. Gottesman, I. Chuang, *Nature* **402**, 390 (1999).
4. A. Barenco *et al.*, *Phys. Rev. A* **52**, 3457 (1995).
5. A. Y. Kitaev, *Russ. Math. Surv.* **52**, 1991 (1997).
6. P. W. Shor, *SIAM J. Comput.* **26**, 1484 (1997).
7. R. P. Feynman, *Int. J. Theoret. Phys.* **21**, 467 (1982).
8. S. Lloyd, *Science* **273**, 1073 (1996).
9. L. K. Grover, *Phys. Rev. Lett.* **79**, 325 (1997).
10. I. L. Chuang, N. Gershenfeld, M. Kubinec, *Phys. Rev. Lett.* **80**, 3408 (1998).
11. A. R. Calderbank, P. W. Shor, *Phys. Rev. A* **54**, 1098 (1996).
12. A. M. Steane, *Phys. Rev. Lett.* **77**, 793 (1996).
13. J. Kim, O. Benson, Y. Yamamoto, *Nature* **397**, 500 (1999).
14. P. Michler *et al.*, *Science* **290**, 2282 (2000).
15. C. H. Bennett, G. Brassard, *Proceedings of IEEE International Conference on Computers, Systems, and Signal Processing*, "Quantum Cryptography: Public Key Distribution and Coin Tossing," Bangalore, India, December 1984 (IEEE, New York, 1984), pp. 175–179.
16. J. I. Cirac, P. Zoller, *Phys. Rev. Lett.* **74**, 4091 (1995).
17. C. E. Wieman, D. E. Pritchard, D. J. Wineland, *Rev. Mod. Phys.* **71**, S253 (1999).
18. J. Ye, D. W. Vernooy, H. J. Kimble, *Phys. Rev. Lett.* **83**, 4987 (1999).
19. B. E. Kane, *Nature* **393**, 133 (1998).
20. N. A. Gershenfeld, I. L. Chuang, *Science* **275**, 350 (1997).
21. D. G. Cory, A. F. Fahmy, T. F. Havel, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 1634 (1997).
22. D. Loss, D. P. DiVincenzo, *Phys. Rev. A* **57**, 120 (1998).
23. J. E. Mooij *et al.*, *Science* **285**, 1036 (1999).
24. N. Linden, H. Barjat, E. Kupce, R. Freeman, *Chem. Phys. Lett.* **307**, 198 (1999).

VIEWPOINT

The World-Wide Telescope

Alexander Szalay,[1] Jim Gray[2]

All astronomy data and literature will soon be online and accessible via the Internet. The community is building the Virtual Observatory, an organization of this worldwide data into a coherent whole that can be accessed by anyone, in any form, from anywhere. The resulting system will dramatically improve our ability to do multi-spectral and temporal studies that integrate data from multiple instruments. The Virtual Observatory data also provide a wonderful base for teaching astronomy, scientific discovery, and computational science.

Many fields are now coping with a rapidly mounting problem: how to organize, use, and make sense of the enormous amounts of data generated by today's instruments and experiments. The data should be accessible to scientists and educators so that the gap between cutting-edge research and education and public knowledge is minimized and should be presented in a form that will facilitate integrative research. This problem is becoming particularly acute in many fields, notably genomics, neuroscience, and astrophysics. The availability of the Internet is allowing new ideas and concepts for data sharing and use. Here we describe a plan to develop an Internet data resource in astronomy to help address this problem in which, because of the nature of the data and analyses required of them, the data remain widely distributed rather than gathered in one or a few databases (e.g., GenBank). This approach may be applicable to many other fields. Our goal is to make the Internet act as the world's best telescope—a World-Wide Telescope.

Today, there are many impressive archives painstakingly constructed from observations associated with an instrument. The Hubble Space Telescope (HST) (*1*), the Chandra X-Ray Observatory (*2*), the Sloan Digital Sky Survey (SDSS) (*3*), the Two Micron All Sky Survey (2MASS) (*4*), and the Digitized Palomar Observatory Sky Survey (DPOSS) (*5*) are examples of this. Each of these archives is interesting in itself, but temporal and multi-spectral studies require combining data from multiple instruments. Furthermore, yearly advances in electronics bring new instruments, doubling the amount of data we collect each year (Fig. 1). For example, approximately a gigapixel is deployed on all telescopes today, and new gigapixel instruments are under construction. A night's observation requires a few hundred gigabytes of memory. The processed data for a single spectral band over the whole sky, a few terabytes. It is impossible for each astronomer to have a private copy of all the data they use. Many of these new instruments are being used for systematic surveys of our galaxy and of the distant universe. Together they will give us an unprecedented catalog to study the evolving universe, provided that the data can be systematically studied in an integrated fashion.

Online archives already contain raw and derived astronomical observations of billions of objects from both temporal and multi-spectral surveys. Together, they house an order of magnitude more data than any single instrument. In addition, all the astronomy literature is online and is cross-indexed with the observations (*6*, *7*).

Why is it necessary to study the sky in such detail? Celestial objects radiate energy over an extremely wide range of wavelengths from radio waves to infrared, optical to ultraviolet, x-rays and even gamma rays. Each of these observations carries important information about the nature of the objects. The same physical object can appear to be totally different in different wavebands (Fig. 2). A young spiral galaxy appears as many concentrated "blobs," the so-called HII regions in the ultraviolet, whereas in the optical it appears as smooth spiral arms. A galaxy cluster can only be seen as an aggregation of galaxies in the optical, whereas x-ray observations show the hot and diffuse gas between the galaxies.

The physical processes inside these objects can only be understood by combining observations at several wavelengths. Today, we already have large sky coverage in 10 spectral regions; soon we will have additional data in at least five more bands. These will reside in different archives, making their integration all the more complicated.

Raw astronomy data is complex. It can be in the form of fluxes measured in finite size pixels on the sky, spectra (flux as a function of wavelength), individual photon events, or

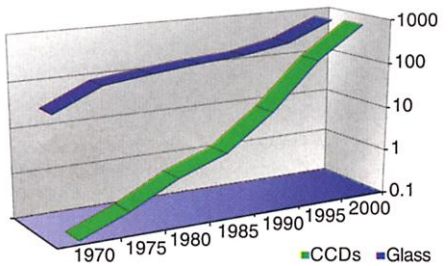

Fig. 1. Telescope area doubles every 25 years, whereas telescope CCD pixels double every 2 years. This rate seems to be accelerating. It implies a yearly data doubling. Huge advances in storage, computing, and communications technologies have enabled the Internet and will enable the Virtual Observatory.

[1]The Johns Hopkins University, Baltimore, MD 21218, USA. [2]Microsoft Bay Area Research Center, San Francisco, CA, USA.

even phase information from the interference of radio waves.

In many other disciplines, once data is collected, it can be frozen and distributed to other locations. This is not the case for astronomy. Astronomy data needs to be calibrated for the transmission of the atmosphere and for the response of the instruments. This requires an exquisite understanding of all the properties of the whole system, which sometimes takes several years. With each new understanding of how corrections should be made, the data are reprocessed and recalibrated. As a result, data in astronomy stays "live" much longer than in other disciplines—it needs an active "curation," mostly by the expert group that collected the data.

Consequently, astronomy data reside at many different geographical locations, and that will not change. There will not be a central "Astronomy database." Each group has its own historical reasons to archive the data one way or another. Any solution that tries to federate the astronomy data sets must start with the premise that this trend is not going to change substantially in the near future; there is no top-down way to simultaneously rebuild all data sources.

To solve these problems, the astrophysical community is developing the World-Wide Telescope, often called the "Virtual Observatory" (*8*). In this approach, the data will primarily be accessed via digital archives that are widely distributed. The actual telescopes will either be dedicated to surveys that feed the archives, or telescopes will be scheduled to follow up on "interesting" phenomena found in the archives. Astronomers will look for patterns in the data—spectral and temporal, known and unknown—and use these to study various object classes. They will have a variety of tools at their fingertips: a unified search engine, to collect and aggregate data from several large archives simultaneously, and a huge distributed computing resource, to perform the analyses close to the data, in order to avoid moving petabytes of data across the networks.

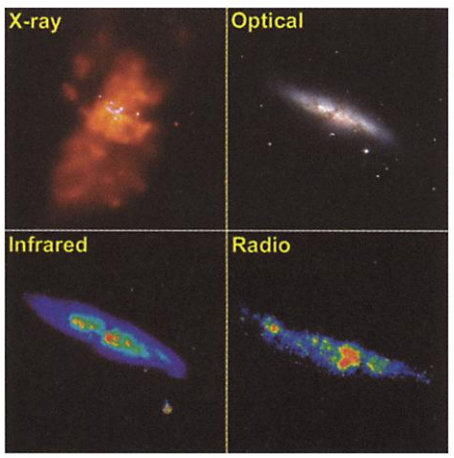

Fig. 2. M82, at a distance of 11 million light years, is a rare nearby starburst galaxy where stars are forming and expiring at a rate 10 times higher than in our galaxy. A close encounter with the large galaxy M81 in the last 100 million years is thought to be the cause of this activity. The images taken at different wavelengths unveil different aspects of the star-forming activity inside the galaxy. Images courtesy of: NASA/CXC/SAO/PSU/CMU (x-ray), AURA/NOAO/NSF (optical), SAO (infrared), and MERLIN/VLA (radio). Compilation from http://chandra.harvard.edu/photo/0094/index.html.

Other sciences have comparable efforts of putting all their data online and in the public domain—GenBank in genomics is a good example—but so far these are centralized rather than federated systems.

The Virtual Observatory will give everyone access to data that span the entire spectrum, the entire sky, all historical observations, and all the literature. For publications, data will reside at a few sites maintained by the publishers. These archive sites will support simple searches. More complex analyses will be done with imported data extracts at the user's facility.

Time on the instrument will be available to all. Thus, the Virtual Observatory should make it easy to conduct such temporal and multispectral studies by automating the discovery and the assembly of the necessary data.

One of the main uses of the Virtual Observatory will be to facilitate searches where statistics are critical. We need large samples of galaxies in order to understand the fine details of the expanding universe and of galaxy formation. These statistical studies require multicolor imaging of millions of galaxies and measurement of their distances. We need to perform statistical analyses as a function of their observed type, environment, and distance.

Other projects study rare objects, ones that do not fit typical patterns; they search for the needles in the haystack. To this end, the use of multi-spectral observations is an enormous help. Colors of objects reflect their temperature. And in the expanding Universe, the light emitted by distant objects is redshifted. Therefore, searching for extremely red objects finds either extremely cold objects or extremely distant ones. Data mining studies of extremely red objects discovered distant quasars, the latest at a redshift of 6.28 (*9*). Mining the 2MASS and SDSS archives found many cold objects such as brown dwarfs, which are bigger than a planet yet smaller than a star. These are good examples of multiwavelength searches not possible with a single observation of the sky, done by hand today, automated in the future. We do not even know all of the data that existed; we will have to discover them on the fly.

The Time Dimension

Most celestial objects are essentially static; the characteristic timescale for variations in their light output is measured in millions or billions of years. There are time-varying phenomena on much shorter timescales as well. Variations are either transient, like supernovae, or regular, like variable stars. If a dark object in our galaxy passes in front of a star or galaxy, we can measure a sudden brightening of the background object, due to gravitational microlensing. Asteroids can be recognized by their rapid motion. All these variations can happen on the scale of a few days. Stars of the Milky Way Galaxy are all moving in its gravitational field. Although few stars can be seen to move in the matter of days, comparing observations 10 years apart accurately measures such motions.

Identifying and following object variability is time consuming and adds an additional dimension to the observations. Not only do we need to map the Universe at many different wavelengths, we need to do it often, so that we can find the temporal variations on many time scales. Once this ambitious gathering of possibly petabyte-size data sets is under way, we will need summaries of light curves; we will also need extremely rapid triggers. For example, in gamma-ray bursts, much of the action happens within seconds after the burst is detected. This puts stringent demands on data archive performance.

The architecture must be designed with a 50-year horizon, because things change significantly in that time—computers will be several orders of magnitude faster, cheaper, and smarter. So the architecture must not make short-term technology compromises. On the other hand, the system must work today on today's technology.

The Virtual Observatory will be a federation of astronomy archives, each with unique assets. Typically, archives will be associated with the institutions that gathered the data and with the people who best understand the data. Some archives might contain data derived from others, and some might contain synthetic data from simulations. A few archives might specialize in organizing the astrophysical literature and cross-indexing it with the data, while others might just be indices of the data itself, analogous to Yahoo! for the text-based Web.

Astronomers own the data they collect, but the field has a long tradition of making all data public after a year. This gives the astronomer time to analyze data and publish early results, and it also gives other astronomers timely access to the data. Given that data are doubling every year and that the data become public within a year, about half the world's astronomy data is available to all. A few astronomers have access to a private data stream from some instrument, so we estimate that everyone has access to 50% of the data and some have access to 55% of the data.

Uniform Views of Diverse Data

The social dynamics of the Virtual Observatory will always have a tension between coherence and creativity, or between uniformity

and autonomy. It is our hope that the Virtual Observatory will act as a catalyst to homogenize the data. It will constantly struggle both with the diversity of the different collections and the creativity of scientists who want to innovate and discover new concepts and new ways of looking at things. These two forces need to be balanced.

Each individual archive will be an autonomous unit run by scientists. The challenge is to translate this heterogeneous mix of data sources into a uniform body of knowledge for the scientists and educators who want to use data from multiple sources. Each archive needs to easily present its data in compatible formats, and the archives must be able to exchange data.

This uniform view will require agreement on terminology, on units, and on representations—a unified conceptual schema (data model) that describes all the data in all the archives with a common terminology, like Vizier (*10*). This schema will evolve with time, and there will always be things that are outside the schema, but Virtual Observatory users will see all the archives via this unifying schema, and data interchange will be done in terms of the schema.

We believe that the base representations will likely be done using the emerging standards for XML, Schemas, SOAP, and Web services (*11*), but there will have to be tools beyond these that automatically transform the diverse and heterogeneous archives to these common formats. Achieving this is beyond the current state of computer science, but solving this schema integration problem will be key for the Virtual Observatory.

Users will want to query the Virtual Observatory using graphical tools, both to pose questions and to analyze and visualize the results. The users will range in skill from professional astronomers to bright grammar school students, so a variety of tools will be needed.

The Virtual Observatory query systems will need to be able to find data among the versions and replicas stored at the various archives. The user might ask for the correlation between radio, optical, and x-ray sources with certain properties. It will be up to the query system to locate the appropriate data sets, subset them, and cross-correlate them, returning the requested attributes from each data source. As with schema integration, there is good technology for automatic indexing, location-transparent, and parallel data search. But the scenario just described is beyond the limits of what computer science researchers can do today.

At its core, the Virtual Observatory will be a data manager. In astronomy, there is a well-established discipline of defining and managing data. There are good languages for defining data schemas (structures) and for mapping these data schemas into existing data management systems. But most of these tools are designed for homogeneous environments with central control, for example, managing the assets of an individual corporation, organization, or agency. Tools that integrate heterogeneous databases are still primitive and labor intensive.

The raw astronomy data from a telescope runs through a substantial "software pipeline" that extracts objects (stars, galaxies, clouds, planets, asteroids) from the data and assigns attributes (luminosity, morphology, classification) to them. These pipelines are constantly improving as we learn new properties of astronomy and as we discover flaws in the pipeline software, so the derived data is constantly evolving into more current versions.

Each archive's data will likely be replicated at one or more mirror sites so that a catastrophe at one site will not cause any long-term data loss. Data may also be wholly or partially replicated at other sites so to speed local computations, because it is often faster and cheaper to access local data than to fetch it from a remote site.

We assume the Internet will evolve so that copying larger data sets will be feasible and economic (in contrast, today data sets in the terabyte range are typically moved by parcel post). Still, it will often be best to move the computation to petabyte-scale data sets in order to minimize data movement and speed the computation.

Current data management systems propagate and manage data replicas, but they require administrators to decide when and where the replicas are stored. Administrators must design and manage the replication strategy and must track data versions and data lineage. As the programs evolve, the computer science challenge is to automate these tasks of configuring and tracking replicas and of tracking data lineage and dependencies on different derived data versions.

Better Algorithms

At the micro level, we expect major advances in computer science algorithms. Hardware performance has improved about 100-fold every decade since 1970. There has been a comparable improvement in software algorithms, the simplest of which is sorting: sort performance has doubled each year since 1985, partly to improvements in hardware and partly to improvements in parallel sorting algorithms. We must continue to invest in and investigate new algorithms and data structures for data loading, cleansing, searching, organization, and mining. Statistical methods lie at that heart of most of our data search algorithms, so advances in computational statistics will be essential to the success of the Virtual Observatory.

Current computer architectures are moving toward huge arrays of loosely coupled computers with local storage. These *Beowulf* clusters harness commodity components and provide very inexpensive computing. Cellular architectures like IBM's BlueGene project or the Japanese Earth Simulator project carry this to an extreme, millions of computers in one cluster. Computational Grids are clusters of computer clusters that can dynamically allocate resources to the tasks at hand (*12*). However, not all our algorithms fit this computational structure; some require fine-grain parallelism and shared memory. We must either find new algorithms that work well on cluster computer architectures or invest in massive-memory computer machines that can solve these problems.

Education

The Virtual Observatory offers the opportunity to teach science in a participatory way. We can give students direct access to a wonderful scientific instrument. They can use it to make discoveries on their own. Very interesting projects and lectures can be built using the Virtual Observatory tools and data.

Astronomy holds a particular attraction for many students, adults and children alike, as demonstrated by planetariums, amateur telescopes, and textbooks. Even very young children can be engaged in many different sciences via astronomy with its strong ties to physics, chemistry, and mathematics. Astronomy can be used as a vehicle for introducing the basic concepts of all these fields and also used to teach the process of scientific discovery. We, the authors, with a lot of help from others, are in the process of creating such a pilot project using data from the SDSS (*13*).

The Virtual Observatory can also be used to teach computational science. Traditionally, science has been either theoretical or empirical. In the past 50 years, computational science emerged as a third approach, first with simulations and now with mining scientific data. Indeed, most scientific departments now have a strong computational program. The Virtual Observatory is a unique tool to teach these skills.

As with astronomy, the Virtual Observatory is a worldwide effort. Initiatives are underway in many countries with a common goal: to join the diverse worldwide astronomical databases into a single federated entity facilitating new research for the astronomy community. The European national and international astronomy data centers are leading an effort funded by the European Union. The Astronomical Virtual Observatory (AVO) is led by the European Southern Observatory and also backed by the European Space Agency. The European Union also sponsors research in pipeline processing technology to handle the anticipated terabyte data streams from future large survey telescopes. The United Kingdom funds the AstroGrid Project to investigate distributed data archives Grid technology. Japan and Australia are setting up their own large archives. In the United States, the National Science Foundation

sponsors development of the information technology infrastructure necessary for a National Virtual Observatory (NVO). There is a close cooperation with the particle physics community through the Grid Physics Network (GriPhyN). NASA supports astronomy mission archives and discipline data centers while developing a roadmap for their federation.

Impressively, these projects are all cooperating, and are working toward a future Global Virtual Observatory to benefit the international astronomical community and the public alike. There are similar efforts under way in other areas of science as well. The Virtual Observatory has had and will have significant interactions with other science communities, both learning from some and providing a model for others.

References

1. The Hubble Space Telescope, www.stsci.edu.
2. Chandra X-Ray Observatory Center, http://chandra.harvard.edu.
3. The Sloan Digital Sky Survey website, www.sdss.org.
4. The Two Micron All Sky Survey, www.ipac.caltech.edu/2mass.
5. Digitized Palomar Observatory Sky Survey, www.astro.caltech.edu/~george/dposs.
6. SIMBAD Astronomical Database, http://simbad.u-strasbg.fr.
7. NASA/IPAC Extragalactic Database, http://nedwww.ipac.caltech.edu.
8. *Virtual Observatories of the Future*, American Society for Physics Conference Series, vol. 25, R. J. Brunner, S. G. Djorgovski, A. S. Szalay, Eds., (The Astronomical Society of the Pacific, San Francisco, 2001).
9. X. Fan *et al.*, e-Print available at http://xxx.lanl.gov/abs/astro-ph/0108063.
10. Vizier Service, http://vizier.u-strasbg.fr/viz-bin/VizieR/
11. The Semantic Web, www.w3.org/2001/sw/; Web Services, www.w3.org/TR/wsdl.
12. I. Foster, C. Kesselman, Eds., *The Grid: Blueprint for a New Computing Infrastructure* (Kaufmann, San Francisco, 1998).
13. Public access Web site to the SDSS data, http://skyserver.sdss.org/

VIEWPOINT

Pathway Databases: A Case Study in Computational Symbolic Theories

Peter D. Karp

A pathway database (DB) is a DB that describes biochemical pathways, reactions, and enzymes. The EcoCyc pathway DB (see http://ecocyc.org) describes the metabolic, transport, and genetic-regulatory networks of *Escherichia coli*. EcoCyc is an example of a computational symbolic theory, which is a DB that structures a scientific theory within a formal ontology so that it is available for computational analysis. It is argued that by encoding scientific theories in symbolic form, we open new realms of analysis and understanding for theories that would otherwise be too large and complex for scientists to reason with effectively.

What happens when a scientific theory is too large to be grasped by a single mind? Decades of experimentation by molecular biologists to characterize the molecular components of single cells, combined with recent advances in genomics, have thrust biology into the position where the theoretical understanding of a system such as the biochemical network of *E. coli* is too large for a single scientist to grasp. This situation has a number of disturbing consequences: It becomes extremely difficult to determine whether such theories are internally consistent or are consistent with external data, to refine theories that are inconsistent, or to understand all of the implications of such large theories. As more details of such a complex system are elucidated experimentally, it is not so clear that our understanding of the system as a whole increases if the new understanding cannot be integrated with the larger theory it pertains to in a coherent fashion.

In this article I argue that as scientific theories reach a certain complexity, it becomes essential to encode those theories in a symbolic form within a computer database (DB). I describe pathway DBs as a case study in encoding scientific theories in computers. Although the scientific community clearly accepts the need to encode the ever-expanding quantity of scientific data within DBs, DBs of scientific theories, such as a theory describing the transcriptional regulation of *E. coli* genes, are much rarer. By data I mean measurements made from individual experiments; by theory I mean relationships inferred from the interpretation and synthesis of many experimental results. The biological sciences are particularly well suited to the DB approach because many theories in biology have a qualitative nature; they describe semantic relationships between systems with many different molecular components, and the causal relationships between these components have been measured in a qualitative rather than a quantitative fashion. The DB approach is probably less appropriate for quantitative theories that are best described by systems of differential equations, or other types of mathematical models in analytical form.

The theory of the *E. coli* metabolic network is an example of a theory whose size and complexity are too large for a mind to grasp. The metabolic network is essentially a chemical processing factory within each *E. coli* cell that enables the organism to convert small molecule chemicals that it finds in its environment into the building blocks of its own structures, and to extract energy from those chemicals. The *E. coli* metabolic network, illustrated in Fig. 1, involves 791 chemical compounds organized into 744 enzyme-catalyzed biochemical reactions (*1*). On average, each compound is involved in 2.1 reactions. I posit that the majority of scientists cannot grasp every intricate detail of this complex network. Omission of even a single step from the network can be fatal for the cell.

One might argue that the biomedical literature is one embodiment of the theory of the *E. coli* metabolic network, and that as the biomedical literature enters electronic form, we need not be concerned with the size and complexity of biology theories. Although efforts to bring the biomedical literature online are tremendously useful, there are serious limitations to what they will achieve: We cannot compute effectively with theories within the biomedical literature. Natural-language texts still remain largely opaque to computers, despite many advances in natural-language processing. For example, one relatively simple question we might wish to ask of the *E. coli* metabolic network is how many of its reaction steps are catalyzed by multiple enzymes, meaning they have backup systems, and therefore would targeting a drug toward one of the enzymes catalyzing those steps be ineffective? Answering this question by using a pathway DB such as the EcoCyc pathway DB is trivial, but answering this question by processing the biomedical literature with a computer program would earn the programmer a Ph.D. in computer science.

Pathway Databases

A pathway is a linked set of biochemical reactions—linked in the sense that the product of one reaction is a reactant of, or an enzyme that catalyzes, a subsequent reaction. A pathway DB is a bioinformatics DB that describes biochemical pathways and their component reactions,

Bioinformatics Research Group, SRI International, EK223, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA. E-mail: pkarp@ai.sri.com