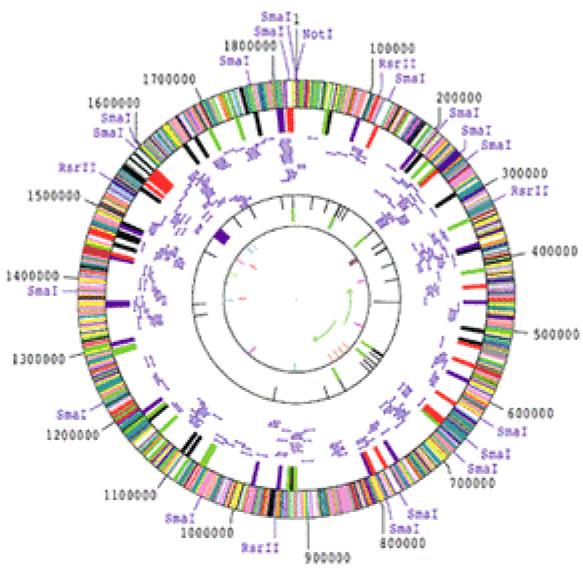
Oscar Gruss

BIOTECHNOLOGY REVIEW

13 March 2000

Jason Reed, PhD.
(212) 514-2341
jreed@oscargruss.com

Trends in Commercial Bioinformatics



Genome of *H. influenzae (TIGR)*

KEY POINTS:

- Roughly defined, "bioinformatics technology" is the backbone computational tools and databases that support genomic and related research.
- The spectacular rise of the commercial genomics industry and the broadening application of genomic techniques in biology and medicine has created a commercial market for bioinformatics software, hardware and services.
- By some estimates, the total market for bioinformatics tools and services, including custom databases, could exceed $2.0 billion within five years.
- In our opinion, bioinformatics technology will become an increasingly important competitive differentiator for public and private life science companies going forward.
- Bioinformatics is becoming a directly investible theme. By our estimation, there are now more than 50 companies which offer bioinformatics products and services of various kinds to external customers. Most of these are private companies, but we would not be surprised to see a number of the more mature players go public in the next 12 months.

COMMERCIAL BIOINFORMATICS?

Introduction. The purpose of this document is to provide an overview of the rapidly emerging field of **"commercial bioinformatics."** We assume the reader has at least a baseline understanding of genomic technologies, and how they are now being implemented in commercial drug discovery. For purposes of this review, we define bioinformatics as the backbone computational tools and databases that support genomic and related research, which broadly encompasses the study of DNA structure/function, gene expression and protein production/structure/function. The spectacular rise of the commercial genomics industry and the broadening application of genomic techniques in biology and medicine has created a commercial market for bioinformatics software, hardware and services.

A Flood of DNA Sequence Data. The initiation of large-scale genomic research projects roughly a decade ago engendered an intensive effort to create related information management and analysis tools, largely driven by academic computer scientists associated with the institutions involved. One of the first and most important problems encountered was how to acquire, store and analyze massive amounts of DNA sequence information. Reliable, high-throughput sequencing methods perfected in the past few years are now churning out vast quantities of information --- from complete genomes of several bacteria and archaea (bacteria-like organisms that live in extreme conditions: a third kingdom of life) up to a mostly complete sequence of human chromosome 22, completed in late 1999.

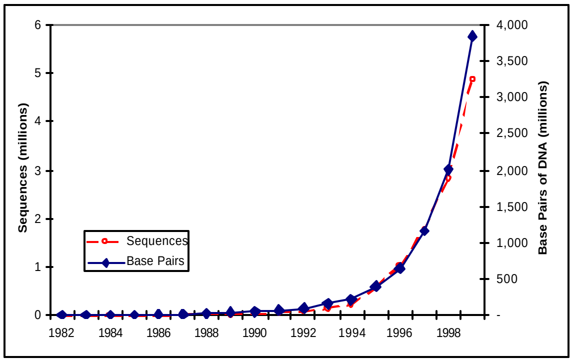
Partial List of Completely Sequenced Genomes

Genome	Size (MM base pairs)	Est. Genes*	Completed	Relevance
Archaea				
Aeropyrum pernix K1	1.67	2,694	1999	Potential source of novel enzymes, etc.
Archaeoglobus fulgidus	2.18	2,407	1997	Potential source of novel enzymes, etc.
Methanobacterium thermoautotrophicum	1.75	1,869	1997	Potential source of novel enzymes, etc.
Pyrococcus abyssi	1.77	1,765	1999	Potential source of novel enzymes, etc.
Pyrococcus horikoshii	1.74	2,064	1998	Potential source of novel enzymes, etc.
Bacteria				
Aquifex aeolicus	1.55	1,522	1997	Potential source of novel enzymes, etc.
Bacillus subtilis	4.21	4,100	1997	Represents sporulating Gram-positive bacteria
Campylobacter jejuni	1.64	1,654	2000	Food-borne pathogen
Chlamydia trachomatis	1.04	894	1998	Human pathogen
Chlamydia pneumoniae	1.23	1,052	1998	Human pathogen
Escherichia coli	4.64	4,289	1998	Key model organism; human pathogen
Haemophilus influenzae	1.83	1,709	1995	Human pathogen; first free-living organism to have genome completely sequenced
Helicobacter pylori	1.67	1,553	1997	Major cause of stomach ulcers
Helicobacter pylori J99	1.64	1,491	1999	Another *H. pylori* strain
Mycobacterium tuberculosis	4.41	3,918	1998	Causes tuberculosis
Mycoplasma genitalium	0.58	480	1995	Genome is interesting because it is very small
Mycoplasma pneumoniae	0.82	677	1996	Leading cause of "walking pneumonia"
Rickettsia prowazekii	1.11	834	1998	Causes epidemic typhus
Synechocystis PCC6803	3.57	3,169	1996	Should help us understand photosynthesis
Treponema pallidum	1.14	1,031	1998	Causes venereal syphilis
Thermotoga maritima	1.86	1,846	1999	Potential source of novel enzymes, etc.
Ureaplasma urealyticum	0.75	611	2000	Sexually transmitted pathogen
Eukaryota				
Caenorhabditis elegans	~97.0	~19,000	1998	Worm – a key model organism
Saccharomyces cerevisiae	12.07	5,885	1996	Yeast – a key model organism
Human Chromosome 22**	33.46	600+	1999	First human chromosome to be fully sequenced

Source: NCBI; *excludes tRNA and rRNA genes; **euchromatic region

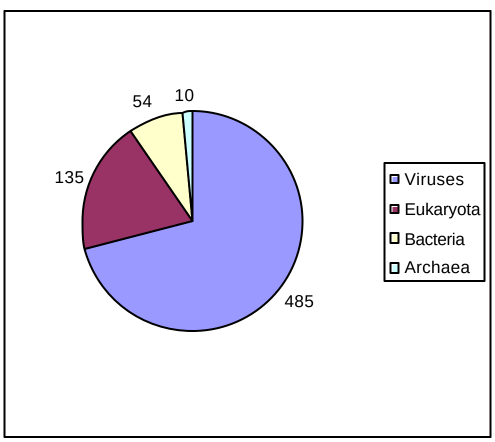
Growth in GenBank. GenBank, a major public repository of DNA sequence data, has grown to include roughly 4.86 million individual sequence records (representing about 3.86 billion base pairs), up from 0.56 million records in 1995 (0.38 billion base pairs). At the time of this writing, GenBank contained the full and partial genome sequences of over 670 different organisms, including 27 complete genomes (6 archaea, 19 bacteria and 2 eukaryotes).

Growth of GenBank



Source: NCBI

Organisms Represented in GenBank



Source: NCBI

This DNA sequence has been deposited in GenBank by a whole host of international academic and government research groups, as well as by commercial concerns. Almost all companies conducting "genomic" research, such as Incyte, Human Genome Sciences, Millennium Pharmaceuticals, Myriad Genetics and Genome Therapeutics, have sequenced stretches of human and other organisms' DNA. Some of this privately-generated sequence data has been submitted to public databases like GenBank, while some remains proprietary.

Human Genome Sequence. A public consortium now plans to produce a draft version of the human genome sequence by mid-2000, and completely sequence the genome by 2003. This effort will be spearheaded by research groups at Washington University (St. Louis), Baylor College of Medicine, the Whitehead Institute and the Sanger Center in England. This Human Genome Project consortium has greatly speeded timelines from its original genome completion date of 2005 due to: (1) the development of robust high throughput sequencing techniques, and (2) competition from Celera, a division of PE Corp that intends to complete a lower fidelity, but still very useful, copy of the human genome by 2001 (with a draft version expected out this year). In addition, both the public consortium and private concerns including Celera are sequencing all or parts of the genomes of "model" organisms, like the mouse, hoping to gain additional insights into genomic structure and function.

Milestones in Human Genome (~3,000 mm bp) Sequencing	
Genetic Map	Complete
Physical Map	Complete
High-Throughput Sequencing Technology	Largely Perfected
Chromosome 22 Sequence	Finished Late 1999
Celera Draft Sequence	Expected Mid 2000
Public Consortium Draft Sequence	90% by Mid 2000
Celera Final Sequence	Expected 2001
Public Consortium Final Sequence	Expected 2003
Other:	
Human Sequence Variation Data	In Progress
Gene Identification	In Progress
Functional Analysis	In Progress
Sequence of Key Model Organisms:	
E.Coli (4.6 mm bp)	Complete
Yeast (12 mm bp)	Complete
C. Elegans (97 mm bp)	Complete
Drosophila (160 mm bp)	Raw Sequence Finished Late 1999 (Celera)
Mouse (2,600 mm bp)	Expected 2002 (Celera)
Rice (400 mm bp)	Expected 2001 (Celera)

Source: NIH, NCBI, Celera

Data Generation is Accelerating. Data generation is only accelerating at this time because: (1) many genomes besides human are being completely sequenced, and (2) high-throughput methods are being perfected in other areas like gene expression assays, protein-protein interaction assays, in vitro or cell-based assays used in drug development and a host of clinically related genetic tests.

Data Source	Drivers
DNA Sequence	High-throughput techniques: --shotgun sequencing --hybrid shotgun/map-based methods --automated capillary electrophoresis --genome maps of various kinds Lots of medically/biologically interesting organisms
Gene Expression Data	Researchers have now found lots of genes: --cDNA sequencing, SAGE, etc. --genomic sequencing w/ gene ID techniques Microarrays can assay thousands of genes at one time (10,000+) Very important for finding/validating drug targets Often involves model organisms
Protein Data	High-throughput techniques: --2-D gels --mass spectrometry --protein-protein interaction assays (yeast 2 & 3 hybrid assays; also can be on chips) --various new in vitro and cell-based assays Structure determination/prediction methods becoming more powerful Very important for finding/validating drug targets Often involves model organisms
Medical Genetics Data	High-throughput techniques: --SNP/polymorphism chip-based assays --SNP/polymorphism mass spec assays --SNP/polymorphism electrophoresis assays Enables tailoring drugs to patients via genetic profile (pharmacogenomics) Enables more efficient patient selection for clinical trials Enables disease predisposition testing

Source: Oscar Gruss Research

Bioinformatics is Becoming Critical to Life Science R&D. If we take the massive generation of biological data as a starting place, bioinformatics technology enables the extraction of information that can be used in commercial drug discovery, clinical diagnostics, agricultural biotechnology and other applications. Currently, this includes three areas: (1) tools that support laboratory experiments; (2) the design, implementation and integration of biological databases; and (3) various analytical tools to determine via computer vs. experiment things like gene location within a chromosome, finding similar genes or proteins from other species and determining the 3-D structure and function of different proteins. These analyses can enable or greatly accelerate drug target identification efforts, drug lead validation and optimization, pharmacogenomic studies and many other biotech applications.

Bioinformatics Technology Involves:
Design, Implementation and Integration of Biological Databases
Aligning Protein and DNA Sequences
Tools That Support Laboratory Experiments
Assembling DNA Sequence Fragments and Creating Genomic Maps
Recognizing and Annotating DNA Sequence Features
Phylogenetic Comparisons
Predicting RNA Secondary Structure
Modeling Protein Structure and Dynamics
These Techniques are Very Useful in R&D Related to:
Commercial Drug Discovery
Improving Clinical Trials
Medical Diagnostics
Pharmacogenomics – Tailoring Medicines to Individuals
Industrial Biotech
Agbiotech

Source: Oscar Gruss Research; Durbin, et al; Altman

A Comment About the Current Limitations of Bioinformatics Technology: A comprehensive assessment of the strengths and weaknesses of the different bioinformatics technologies is beyond the scope of this report. Suffice it to say that many of these methods are still very much "in development." However, these tools can be very powerful when applied in the correct analytical context, and in conjunction with the appropriate experimental validation. To further clarify how these tools are used, we provide a simplified example below.

BIOINFORMATICS TECHNOLOGY: AN EXAMPLE

Sequence an Interesting Genomic Region. You might start by finding the DNA sequence of a chromosomal region, speculating that it contains genes from an interesting biological pathway. Your ultimate goal might be to find an undiscovered drug target. To rapidly assemble a contiguous DNA sequence that might have one or more complete genes you might use the "shotgun" technique. This technique relies on piecing together many small, electrophoretically determined stretches of DNA sequence, each say 500 base pairs in length, into a much larger continuous stretch, say 2 million base pairs in length. To do this in a (mostly) automated fashion, you will need special programs like PHRED to read the raw DNA sequence, and PHRAP to assemble the small pieces into a large stretch of sequence. You will probably also need to use a laboratory information management system (LIMS) to track your sequencing project, as the process involves many individual samples and pieces of data that need to be stored and organized.

Find Genes and Other Interesting Features in Your Genomic Sequence. You now have a DNA sequence that is a string of several million symbols (like ...AAGGCTGAGTGCTAAGCGC GCG…), or a few strings of several hundred thousand symbols if you cannot put it all together (a common problem). You want to find regions that correspond to genes and perhaps regulatory sequences that control when the genes are turned on and off. You might start by using a program called BLAST (Basic Local Alignment Search Tool) to search the public or commercial DNA sequence databases to see if any stretches of your 2 million base pair sequence match previously identified gene sequences. To do this faster you might use special computer hardware known as an "accelerator," such as the DeCypher system from TimeLogic. You might use a more sophisticated software package like GENIE, GENSCAN or GRAIL to better identify where in your sequence the gene starts and stops, and where regulatory regions might be. These "gene finding" programs are not completely reliable in most cases, but are useful when used in conjunction with other methods. In this regard, you would also want to search public and private expressed sequence tag (EST) databases. ESTs are short sequences (several hundred base pairs) experimentally determined to correspond to real genes. If an EST matches part of your sequence, it is likely that that part contains a real gene. This is a very powerful technique, as you don't actually have to know what the gene does, and because available EST databases are now very comprehensive.

Compare Your Gene to Other Known Genes to Find More Information. As an extension of the process described above, you might want to keep comparing regions of your sequence that you now think correspond to a gene with other known genes from human and with genes from other organisms. If your gene is similar to a human gene of known function, say an enzyme of some kind, your gene might perform the same function and be structurally similar. To do this, you want to continue to use "pairwise alignment" algorithms (like BLAST and Smith-Waterman) to search both public and private databases. Comparing your gene to similar genes in other organisms (say genes from a mouse and a fish, if they are known) can help you find important regulatory and functional regions, among other things, because these tend to be evolutionarily conserved. This is one reason that the Human Genome Project Consortium and Celera are sequencing model organisms like mouse and fruit fly, in addition to human. To compare the sequences of several genes, you can use "multiple alignment" algorithms such as CLUSTAL W, or MSA. There are many other tools to compare the DNA and protein sequence of your new gene with other known genes and protein motifs.

Displaying this Information is Critical to Understanding It. As you have collected all of this information about your DNA sequence, where the genes are, where the regulatory sequences might be, what corresponds to an EST sequence, etc., you will need to display it graphically. One of the best ways to do this is in a "browser" fashion that lets you easily investigate each piece of information via mouse click or something similar. A good display can tell you what information might be lacking and where the different sources of information agree or disagree.

Next Steps. In the example above, we might have been able to find out a great deal about the function, structure and pathway of action of our gene via computer tools. This might tell us that the gene produces a protein that could be important in a disease process. Therefore we have gone from information about a chromosomal region to a potential drug target. **This information can help us more efficiently design future experiments, or make some experiments unnecessary.**

Going forward, we might want to use microarrays to investigate the expression of our gene under different conditions (in response to different chemicals, etc.). There is now great interest in making databases of this type of gene expression information, so you might not have to conduct the experiment yourself. Examples include the gene expression data available commercially from GeneLogic and Incyte, and a host of academic and government research groups now developing free gene expression databases like Stanford University, the Whitehead Institute and the U.S. National Center for Biotechnology Information (NCBI, part of the NIH).

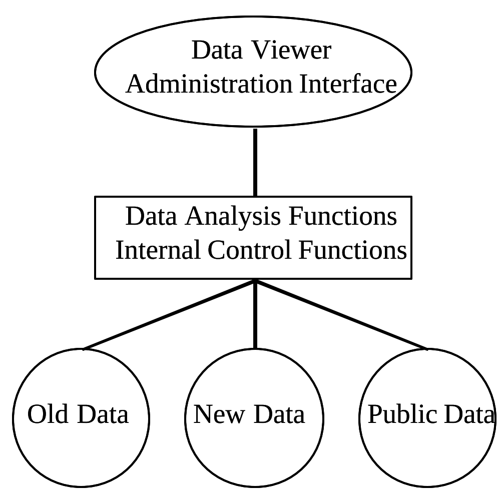
For the purposes of drug development, you'd definitely be interested in which other proteins interact with the protein from your gene. It may turn out that another, structurally dissimilar protein in the pathway would be a better drug target for some reason. (Myriad's ProNet technology, or CuraGen's protein-protein interaction databases are powerful, commercially available tools to find this type of information.)

Finally, you might attempt to model the structure of your protein, and how it interacts with a drug molecule. This might tell you which chemical class of molecules would be the most promising drug candidates. It should be noted that modeling 3-D protein structures and protein-small molecule interactions are some of the toughest problems in computational biology. Companies like Structural Bioinformatics and IBM are working on these kinds of problems, as are many other commercial and academic groups.

With this admittedly oversimplified example, we've hoped to demonstrate the following points about bioinformatics technology:

1. In one form or another, it is ubiquitous in genomics research.
2. It can involve lots of database searching. The more high-quality information available, the more powerful bioinformatics can become.
3. It can require many different algorithms and analyses.
4. Integrating and displaying the information is key.
5. **Bioinformatics won't replace experiments, but can greatly streamline and enable the discovery process.**

An integrated system comprised of databases, analysis algorithms and display tools is described in the schematic below.



Integrating these elements in an easily navigable system, whether it be a desktop program or an enterprise wide IT system, is highly desirable in most commercial and in many large-scale academic research efforts.

Commercial Applications of Bioinformatics Are Numerous. To understand the commercial relevance of these technologies, one need only consider that all of the public and private sector genomics research now being conducted relies heavily on bioinformatics tools of the kind described above. In the future, bioinformatics tools should see extensive use in all the key life science R&D markets, including the pharmaceutical and biotechnology industries, agricultural biotechnology, and in government and academic research. Penetration of bioinformatics techniques into these markets should be driven by the following factors:

1. The pressure to rapidly organize, integrate and mine data is enormous: it costs a lot to produce, and competitive and patent concerns are an issue.
2. Maturation of tools should make them easier to use.
3. Life sciences R&D organizations are becoming more receptive to a paradigm shift in research techniques (i.e. genomics, R&D outsourcing, etc.), due in large measure to the insufficiency of current methods (product output per research $).

But partly offset by:

1. The fact that experienced bioinformatics people are relatively scarce.
2. Lack of universal compatibility standards for tools and databases.
3. Applications can be very complex and heterogeneous, thus the development time/cost is often high.
4. In some cases, the capital expenditure to support in-house capability is quite large, plus constant service expenditures of some type will probably be required.

THE COMMERCIAL BIOINFORMATICS MARKET

Market Structure

By our estimation, there are now more than 50 companies that offer bioinformatics products and services of various kinds to external customers. In surveying the industry, we find surging volume growth, particularly among industry leaders, and an acceleration in the number of corporate deals and other collaborations. We believe this reflects the explosive growth in genomic and related research techniques, plus the weaknesses of available analysis tools and databases. By some estimates, the total market for bioinformatics tools and services, including custom databases, could exceed $2.0 billion within five years.

There remain a number of **significant challenges** in this market, however. Over the past few years, the customer base willing to pay big dollars for a customized bioinformatics solution, large biopharma, has been relatively concentrated (perhaps fewer than 50 customers) and the largest players have mostly satisfied their own needs with in-house bioinformatics expertise. Further, publicly available tools and databases are ubiquitous and becoming easier to use and are more integrated. Commercial solutions that add substantial value tend to be complex with longer development cycles than traditional software products. On the other hand, the individual applications can be very heterogeneous, so it can be hard to leverage a specific product across many applications. The net result is that development time/cost can be high, but each individual market can be relatively narrow. Notably, the recent dissolution of the high-profile bioinformatics startup Molecular Applications Group (MAG) can be traced to these issues. Some bioinformatics companies have responded to these hurdles by reorienting their business models. For example, Pangea Systems recently changed from being an enterprise IT solution provider (a low-volume, high-price business) to being an e-bioinformatics portal (now called DoubleTwist.com), which is targeted mostly toward small- and mid-size customers (high-volume, low-price). Compugen, which originally produced special computer hardware for DNA and protein sequence analysis, now offers expanded services such as DNA microarray design and an e-bioinformatics portal (called LabOnWeb.com).

Because bioinformatics is becoming such a critical enabling technology in modern biological research, we strongly feel that commercial solutions will ultimately reach their multi-billion dollar sales potential. It is an open question as to how the industry will respond to the current problems of market heterogeneity and small customer base. It is our feeling that **consolidation**, driven by the larger players, and cross-platform **standardization** will be major themes going forward.

Below we outline the bioinformatics market structure and growth outlook in further detail:

Product Categories

There are several identifiable bioinformatics product categories: proprietary databases of various kinds, software and hardware analysis tools of varying comprehensiveness, complete enterprise IT systems that manage and integrate databases and analysis tools, and, finally, custom services. In time these distinctions should become blurred as tools, databases and information management systems become more integrated.

We see the following technical hurdles as important to bioinformatics product design, and the solutions which most effectively address them should have a competitive advantage:

1. The data to be organized/analyzed is very heterogeneous.
2. Analysis tools are rapidly evolving.
3. Seamlessly integrating public, legacy and new data is a must.
4. Many users are not software/computer experts.

Customer Base

1. **Pharmaceutical and biotechnology companies** will use bioinformatics technology in all stages of the drug discovery process, from drug target identification through lead validation and optimization to drug response profiling and clinical diagnostics.

   Key driver: This is the most important customer base in terms of dollar value, due to competitive and patent expiry pressures and the fact that biopharma has traditionally spent heavily on R&D. Large pharmaceutical companies are already prodigious customers of outsourced genomics R&D that includes a lot of bioinformatics content. This includes partnerships like those between Millennium Pharmaceuticals and AstraZeneca, Bayer, Pfizer, and Wyeth Ayest for example, or Human Genome Science's deals with SmithKline Beecham, Schering-Plough, Merck KGaA, etc. There are many more examples. We believe that the "middle market" of smaller pharmas and mid- to small-size biotechs (perhaps 300+ companies, excluding genomics companies) is relatively underpenetrated for a variety of reasons, including smaller R&D budgets and a historical emphasis on more traditional drug discovery technologies.

   Key constraint: As discussed above, leading pharma companies that have made a substantial commitment to genomics research have already developed a substantial bioinformatics infrastructure. This includes companies like SmithKline, Glaxo, Merck, Novartis and others. These types of customers are potentially the highest value segment, but displacing a big pharma's custom-tailored bioinformatics group with an external product is only practical in the case of niche or especially high-value applications. Lion Biosciences of Heidelberg, Germany, has been perhaps the most successful bioinformatics company to date in penetrating big pharma with a high-value infrastructure deal. In

1999, Lion entered into a five-year alliance with Bayer AG worth up to $100 million, in which Lion will provide and support bioinformatics IT systems to speed Bayer's drug discovery programs. The deal included the establishment of Lion's U.S. subsidiary Lion Bioscience Research in Cambridge, MA.

2. **Agbiotech/Industrial Biotech companies** have already started to use genomics research methods extensively in the study of crops and livestock, with the hope of improving crop/livestock yields, increasing pesticide/herbicide resistance, improving taste/nutritional content, etc.

   Key driver: We expect that the widening use of gene expression assays and proteomics assays of various kinds in ag/industrial biotech will sharply increase the need for bioinformatics technology in this market. The increased pace of whole genome sequencing of thermophilic organisms and other "extremeophiles" (like that of *M. thermoautotrophicus* by Genome Therapeutics), which may provide a novel source of enzymes for industrial processes, should support this trend.

   Key constraint: This market segment has traditionally been slower than biopharma to embrace genomic techniques. We believe that the current negative public perception of genetically modified organisms (GMOs) will remain a factor, at least in the near future.

3. **Academic research groups**, particularly those associated with the international effort to sequence the human genome, have pioneered most of the genomic and bioinformatics techniques in use today and should continue to be heavy users.

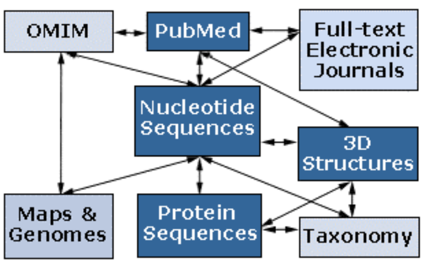
   Key driver: Cutting edge research into gene expression, proteomics and medical genetics will increasingly rely on the use of bioinformatics tools, in our opinion.

   Key constraint: Outside of the large, government coordinated projects like the Human Genome Initiative, individual researchers tend to be less intensive data generators/users than commercial concerns. As a result, their bioinformatics needs are often satisfied by a combination of publicly available tools, commercial desktop solutions (like those available from InforMax or GCG) and home grown systems.

4. **Other markets** include government agencies like the U.S. Patent and Trademark Office, which recently purchased a Compugen DNA and protein sequence analysis computer system to aid in patent searches. We expect law enforcement agencies like the FBI and the armed services to compile and make increasing use of genetic profile databases in the future. However, in the near term, these non-commercial markets will probably remain small in terms of total dollar value.

Participants in the Field

1. **Academic and government groups** which produce publicly available tools and databases, some of which are quite comprehensive and sophisticated. Examples are the many tools and databases maintained by the NCBI, including GenBank. Appendix B at the end of this report contains a partial list of available biological databases, many of which are public free-access databases. Below is a schematic of NCBI's Entrez database browser system:



Source: NCBI

2. **Genomic and pharmacogenomic companies** that offer databases and services to outside customers, as well as for their own internal use. This includes companies like Incyte, Celera, CuraGen and GeneLogic. We would also include biotech instrumentation companies like PE Biosystems in this category. Instrumentation products usually include data management and analysis tools of varying utility.

3. **Large pharma, biotech and agbio companies** which develop their own in-house databases and bioinformatics expertise. As discussed above, some of the largest pharmaceutical companies have well-developed bioinformatics infrastructures, and thus are difficult for outside providers to penetrate. The situation is much more favorable in mid-size to smaller companies, however these firms often cannot justify extremely large expenditures on infrastructure unless it addresses a core research focus.

4. **Traditional computer, electronic technology and IT services companies** that offer products and services for the bioinformatics market. This includes companies like Compaq, Sun Microsystems, Silicon Graphics, IBM and Agilent Technologies. For the most part, these companies have taken the complementary approach of providing infrastucture that supports various solutions by specialized bioinformatics providers. **We expect these companies to be an increasingly important competitive force in genomics and bioinformatics.** For instance, Compaq has a major strategic alliance with Celera to provide integrated bioinformatics hardware, software, networking and service solutions. IBM is conducting research into high value data mining and protein structure determination methods. IBM offers a variety of enterprise-wide IT solutions for the life science market, and recently initiated a collaboration with NetGenics. Through its partnership with Rosetta Inpharmatics, Agilent offers an enterprise-wide gene expression analysis solution that includes software and hardware and is a rival to Affymetrix's GeneChip system.

5. **More or less "pure play" bioinformatics companies** that offer products and services to external customers. Some of these companies are trying to leverage their bioinformatics expertise toward in-house efforts like drug discovery, and are thus somewhat like traditional genomics companies (see category #2 in this list). **Most of these are private companies**, but we would not be surprised to see a number of the more mature players go public in the next 12 months. Some, but by no means all, of the prominent companies in this space are listed in Appendix A of this report, and the market outlook for this segment is discussed in more detail below.

More on Market Size and Growth Outlook

Given the nascent nature of this industry and the large number of private players in the field, the current market for external products and services is hard to determine. Surveys of the 50 or so bioinformatics tool and database companies by market research groups like Frontline and Frost & Sullivan, for example, put the current market for bioinformatics databases, products and services at roughly $300 million, with about half of the annual sales by data suppliers and half of the sales by tool/IT providers of various kinds. These groups and other industry observers believe that this market could grow to $1.5-2.0 billion over the next five years. These estimates exclude some significant internal spending on IT infrastructure by pharmaceutical and biotechnology companies that is bioinformatics related, and could be as large as $2.0+ billion annually. As discussed above, also excluded are most of the project-based R&D collaborations between pharma/agbio companies and genomics companies that include bioinformatics "content," and which total well over $1.0 billion on a cumulative basis over the past 3-5 years.

Without more publicly disclosed financials these market size estimates are hard to pin down. However, we find them reasonable if not conservative, in that they imply visible 25%-35% top-line growth over the next few years, which is consistent with our own survey of key industry players. Conceptually, as discussed above, we believe bioinformatics will become essential to many if not all life science R&D activities, and the market for commercial solutions of various kinds should increase in proportion.

Most of today's sales come from the database providers and the software/hardware tool suppliers, with complete enterprise IT solution just emerging (perhaps 20% or less of the current market). Over the next few years, the enterprise IT solution should garner a larger proportion of the industry's total sales, driven by a great need for integration of the various databases/tools with R&D efforts. Also, we expect growth in the sales of data providers to be supported by the emergence of new types of data, namely gene expression and proteomics data. However, commercial database sales are likely to be constrained by the increasing public availability of well-annotated genome sequence from human and other organisms, and by the increasing public availability of other types of data.

Appendix A -- Representative Bioinformatics Database, Software, Hardware and Service Providers

Concentrated Bioinformatics Plays	Ticker	Description
Compugen	Private	Originally specialized in computer hardware/software designed to accelerate bioinformatics algorithms. Business model now moving more toward an internet portal concept, plus proprietary and collaborative gene discovery.
DoubleTwist.com	Private	An internet portal business model, which includes on-line access to a variety of bioinformatics/biotech tools, databases and other products. DoubleTwist changed its name from Pangea Systems in 1999.
eBioinformatics	Private	Originally a spin-off from the Australian National Genomic Information Service. eBioinformatics provides a variety of web-based bioinformatics tools and databases.
Genomica	Private	Provides enterprise-wide bioinformatics systems and services. Relationships include AstraZeneca, Glaxo Wellcome, Parke Davis and PE Biosystems.
Informax	Private	Desktop and enterprise-wide bioinformatics products. Customer base of over 60 pharma companies, 250 biotechs and 500 universities.
Lion Bioscience	Private	Provides enterprise-wide bioinformatics systems and services. Lion has interest in leveraging technology for proprietary R&D. Lion's $100 MM alliance with Bayer AG largest bioinformatics deal to date.
Molecular Mining	Private	Molecular Mining produces high value-added data mining algorithms than can be used to filter gene expression and other types of data.
Neomorphic	Private	Bioinformatics tools to mine and visualize genomic information. Collaborations with key academic and commercial genomic technology leaders.
Netgenics	Private	Provides enterprise-wide bioinformatics systems and services. Relationships include Pfizer, Abbott, Wyeth Ayerst and IBM.
Oxford Molecular	OMG.LN	Comprehensive business model that includes bioinformatics and related fields of cheminformatics and computational chemistry. In 1997, acquired Genetics Computer Group, maker of the popular Wisconsin desktop bioinformatics product.
Paracel	Private	Specialized computer hardware/software designed to accelerate bioinformatics algorithms. Relationships with many academic and commercial research groups, including PE Corp.
Silicon Genetics	Private	Tools for gene expression analysis and visualization, plus other data-mining applications.
SpotFire	Private	SpotFire offers data visualization software for gene expression as well as products for non-life sciences industries.
Structural Bioinformatics	Private	Bioinformatics tools and databases with a special focus on protein structural information, a critical component of rational drug design
TimeLogic	Private	Specialized computer hardware/software designed to accelerate bioinformatics algorithms. Configurable hardware architecture offers competitive advantage in some cases. Relationships with key academic and commercial research groups, including Stanford University, Roche, Bristol-Myers and Novartis.

Genomic/Biotechnology Companies with Bioinformatics Products	Ticker	Description
Celera	CRA	A division of PE Corp founded to rapidly sequence the human and other genomes, with the intent to supply high value-added genomic data to life science collaborators. Celera has the world's most powerful high-throughput DNA sequencing capability.
CuraGen	CRGN	CuraGen conducts project driven genomic R&D for propriety use and in collaboration with life science partners. CuraGen offers collaborators a variety of well-integrated databases, bioinformatics tools and services.
GeneLogic	GLGC	Offers "GeneExpress" gene expression database products, and other services to the life sciences industry.
Human Genome Sciences	HGSI	HGSI practically founded the commercial genomics industry with its landmark 1993 gene database deal with SmithKline Beecham. HGSI now has collaborations with more than ten commercial partners in areas including gene databases, antibodies, gene therapy and microbial genomics.
Incyte	INCY	A pioneer commercial bioinformatics database company. Provides high-value gene expression, proteomics and other data/analysis tools to pharmaceutical and academic subscribers.
Myriad Genetics	MYGN	Myriad's core competence is therapeutic and diagnostic product development via genomic and proteomic methods. Myriad offers a public version of its high-quality protein interaction database, ProNet, through DoubleTwist.com and through its own Myriad-ProNet.com website.
PE Biosystems	PEB	A division of PE Corp, the premier provider of DNA sequencers and other life science instrumentation. The PE Informatics division offers a variety of software products to life science and other customers.
Rosetta Inpharmatics	Private	Rosetta's core competence is obtaining gene expression and other data in a setting relevant to drug/product discovery for proprietary use and in collaboration with life science partners. Through its commercialization partner Agilent, Rosetta offers an enterprise-wide gene expression analysis solution that includes software and hardware.

Computer, Electronic Technology and IT Services Companies Offering Bioinformatics Products	Ticker	Description
Agilent Technologies	A	In 1999, Agilent entered into a strategic collaboration with Rosetta Inpharmatics to make and sell gene expression analysis systems, including hardware and software.
Compaq	CPQ	Compaq has a major strategic alliance with Celera to provide integrated bioinformatics hardware, software, networking and service solutions.
IBM	IBM	IBM is conducting research into high value-added data mining and protein structure determination methods. IBM offers a variety of enterprise-wide IT solutions for the life science market, and recently initiated a collaboration with NetGenics.
Silicon Graphics	SGI	SGI offers visual computing and high-performance computer systems. SGI systems support a wide variety of bioinformatics software applications.
Sun Microsystems	SUNW	Sun systems support a wide variety of bioinformatics software applications.

Appendix B -- Representative Molecular Biology Databases
(from A. Baxevanis in *Nucleic Acids Research*, 2000, V.28, No.1)

Major Sequence Repositories	
GenBank	All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration
EMBL Nucleotide Sequence Database	All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration
DNA Data Bank of Japan (DDBJ)	All known nucleotide and protein sequences; International Nucleotide Sequence Database Collaboration
Genome Sequence Database (GSDB)	All known nucleotide and protein sequences
TIGR Gene Indices	Non-redundant, gene-oriented clusters
UniGene	Non-redundant, gene-oriented clusters
Comparative Genomics	
Clusters of Orthologous Groups (COG)	Phylogenetic classification of proteins from 21 complete genomes
XREFdb	Cross-referencing of model organism genetics with mammalian phenotypes
Gene Expression	
ASDB	Protein products and expression patterns of alternatively-spliced genes
Axeldb	Gene expression in Xenopus
BodyMap	Human and mouse gene expression data
EpoDB	Genes expressed in vertebrate RBC
FlyView	Drosophila development and genetics
Gene Expression Database (GXD)	Mouse gene expression and genomics
Kidney Development Database	Kidney development and gene expression
MAGEST	Ascidian (Halocynthia roretzi) gene expression patterns
Mouse Atlas and Gene Expression Database	Spatially-mapped gene expression data
PEDB	Normal and aberrant prostate gene expression
Tooth Development Database	Gene expression in dental tissue
TRIPLES	TRansposon-Insertion Phenotypes, Localization and Expression in Saccharomyces
Gene Identification and Structure	
Ares Lab Intron Site	Yeast spliceosomal introns
COMPEL	Composite regulatory elements
CUTG	Codon usage tables
EID	Protein-coding, intron-containing genes
EPD	Eukaryotic POL II promoters
ExInt	Exon-intron structure of eukaryotic genes
IDB/IEDB	Intron sequence and evolution
PLACE	Plant cis-acting regulatory elements
PlantCARE	Plant cis-acting regulatory elements
TransTerm	Codon usage, start and stop signals
TRRD	Regulatory regions of eukaryotic genes
YIDB	Yeast nuclear and mitochondrial intron sequences
Genetic Maps	
GeneMap '99	International Radiation Mapping Consortium human gene map
G3-RH	Stanford G3 and TNG radiation hybrid maps
GB4-RH	Genebridge4 (GB4) human radiation hybrid maps
GDB	Human genes and genomic maps
DRESH	Human cDNA clones homologous to Drosophila mutant genes
GenAtlas	Human genes, markers and phenotypes

HuGeMap	Human genome genetic and physical map data
IXDB	Physical maps of human chromosome X
Radiation Hybrid Database	Radiation hybrid map data
Genomic Databases	
AceDB	Caenorhabditis elegans, Schizosaccharomyces pombe and human sequences and genomic information
FlyBase	Drosophila sequences and genomic information
Mouse Genome Database (MGD)	Mouse genetics and genomics
Saccharomyces Genome Database (SGD)	Saccharomyces cerevisiae genome
AmmtDB	Metazoan mitochondrial DNA sequences
Arabidopsis	Database (AtDB) Arabidopsis thaliana genome
CropNet	Genome mapping in crop plants
CyanoBase	Synechocystis sp. genome
EcoGene	Escherichia coli K-12 sequences
EMGlib	Completely sequenced bacterial genomes and the yeast genome
GOBASE	Organelle genome database
HIV Sequence Database	HIV RNA sequences
Human BAC Ends Database	Non-redundant human BAC end sequences
INE	Rice genetic and physical maps and sequence data
Mendel Database	Database of plant EST and STS sequences annotated with gene family information
MitBASE	Mitochondrial genomes, intra-species variants, and mutants
MitoDat	Mitochondrial proteins (predominantly human)
MITOMAP	Human mitochondrial genome
MITONUC/MITOALN	Nuclear genes coding for mitochondrial proteins
MITOP	Mitochondrial proteins, genes and diseases
Munich Information Center for Protein Sequences (MIPS)	Protein and genomic sequences
NRSub	Bacillus subtilis genome
Phytophthora Genome Initiative Database	Oomycete sequences and genetic maps
RsGDB	Rhodobacter sphaeroides genome
TIGR Microbial Database	Microbial genomes and chromosomes
ZFIN	Zebrafish genetics and development; mutant and wild-type lines
ZmDB	Maize genome database
Intermolecular Interactions	
Database of Ribosomal Crosslinks (DRC)	Ribosomal crosslinking data
DIP	Catalog of protein-protein interactions
DPInteract	Binding sites for Escherichia coli DNA-binding proteins
Metabolic Pathways and Cellular Regulation	
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Metabolic and regulatory pathways
EcoCyc	Escherichia coli K-12 genome, gene products and metabolic pathways
ENZYME	Enzyme nomenclature
EpoDB	Genes expressed during human erythropoiesis
FlyNets	Drosophila melanogaster molecular interactions
Klotho	Collection and categorization of biological compounds
LIGAND	Enzymatic ligands, substrates and reactions
RegulonDB	Escherichia coli pathways and regulation
UM-BBD	Microbial biocatalytic reactions and biodegradation pathways primarily for xenobiotic, chemical compounds
WIT2	Integrated system for functional curation and development of metabolic models

Mutation Databases	
Online Mendelian Inheritance in Man (OMIM)	Catalog of human genetic and genomic disorders
ALFRED	Allele frequencies and DNA polymorphisms
Androgen Receptor Gene Mutations Database	Mutations in the androgen receptor gene
Asthma and Allergy Database	Genetics of allergy and asthma, including linkage studies and mutation data
Asthma Gene Database	Linkage and mutation studies on the genetics of asthma and allergy
Atlas of Genetics and Cytogenetics in Oncology and Hematology	Chromosomal abnormalities in cancer
BTKbase	Mutation registry for X-linked agammaglobulinemia
Cytokine Gene Polymorphism Database	Cytokine gene polymorphisms, in vitro expression and disease-association studies
Database of Germline p53 Mutations	Mutations in human tumor and cell line p53 gene
DbSNP	Single nucleotidepolymorphisms
GRAP Mutant Databases	Mutants of family A G-Protein Coupled Receptors (GRAP)
Haemophilia B Mutation Database	Point mutations, short additions and deletions in the Factor IX gene
HAMSTeRS	Hemophilia A mutation database
HGBASE	Intragenic sequence polymorphisms
HIV-RT	HIV reverse transcriptase and protease sequence variation
Human Gene Mutation Database (HMGD)	Known (published) gene lesions responsible for human inherited disease
Human PAX2 Allelic Variant Database	Mutations in human PAX2 gene
Human PAX6 Allelic Variant Database	Mutations in human PAX6 gene
Human Type I and Type III Collagen Mutation Database	Human type I and type III collagen gene mutations
HvrBase	Primate mtDNA control region sequences
iARC p53 Database	Missense mutations and small deletions in human p53 reported in peer-reviewed literature.
KinMutBase	Disease-causing protein kinase mutations
KMDB	Mutations in human eye disease genes
MmtDB	Mutations and polymorphisms in metazoan mitochondrial DNA sequences
Mutation Spectra Database	Mutations in viral, bacterial, yeast and mammalian genes
NCL Mutations	Mutations and polymorphisms in neuronal ceroid lipofuscinoses (NCL) genes
p53 Databases	Human p53 and hprt mutations; transgenic lacZ and transgenic/bacterial lacI mutations
PAHdb	Mutations at the phenylalanine hydroxylase locus
PMD	Compilation of protein mutant data
RB1 Gene Mutation Database	Mutations in the human retinoblastoma (RB1) gene
Ribosomal RNA Mutational Database	16S and 23S ribosomal RNA mutation database
SV40 Large T-Antigen Mutant Database	Mutations in SV40 large tumor antigen gene
Pathology	
FIMM	Functional molecular immunology data (diseases, antigens, peptides and HLA binding sites
Mouse Tumor Biology Database (MTB)	Mouse tumor names, classification, incidence, pathology, genetic factors
PEDB	Sequences from prostate tissue and cell type-specific cDNA libraries
Protein Databases	
AARSDB	Aminoacyl-tRNA synthetase sequences
DatA	Annotated coding sequences from Arabidopsis
DExH/D	Family Database DEAD-box, DEAH-box and DExH-box proteins
Endogenous GPCR List	G protein-coupled receptors; expression in cell lines
ESTHER	Esterases and [alpha]/[beta] hydrolase enzymes and relatives
FUNPEP	Low-complexity or compositionally-biased protein sequences
GenProtEC	Escherichia coli genes, gene products and homologs

GPCRDB	G protein-coupled receptors
Histone Sequence Database	Histone and histone-fold sequences and structures
HIV Molecular Immunology Database	HIV epitopes
Homeobox Page	Information relevant to homeobox proteins, classification and evolution
Homeodomain Resource	Homeodomain sequences, structures, and related genetic and genomic information
HUGE	Large (>50 kDa) human proteins and cDNA sequences
IMGT	Immunoglobulin, T cell receptor and MHC sequences
InBase	Intervening protein sequences (inteins) and motifs
Kabat Database	Sequences of proteins of immunological interest
LGIC	Ligand-gated ion channel sequences, alignments and phylogeny
Membrane Protein Database	Membrane protein sequences, transmembrane regions and structures
MEROPS	Peptidase sequences and structures
MHCPEP	MHC-binding peptides
NRR	Steroid and thyroid hormone receptor superfamily
Olfactory Receptor Database	Sequences for olfactory receptor-like molecules
OoTFD	Transcription factors and gene expression
Peptaibol	Peptaibol (antibiotic peptide) sequences
PhosphoBase	Protein phosphorylation sites
PKR	Protein kinase sequences, enzymology, genetics, and molecular and structural properties
PPMdb	Arabidopsis plasma membrane protein sequence and expression data
Prolysis	Proteases and natural and synthetic protease inhibitors
PROMISE	Prosthetic centers and metal ions in protein active sites
Protein Information Resource (PIR)	Non-redundant protein sequence database
Receptor Database (RDP)	Receptor protein sequences
Ribonuclease P Database	RNase P sequences, alignments and structures
SENTRA	Sensory signal transduction proteins
SWISS-PROT/TrEMBL	Curated protein sequences
TRANSFAC	Transcription factors and binding sites
Wnt Database	Wnt proteins and phenotypes
Protein Sequence Motifs	
BLOCKS	Protein sequence motifs and alignments
PROSITE	Biologically-significant protein patterns and profiles
Pfam	Multiple sequence alignments and hidden Markov models of common protein domains
O-GLYCBASE	Glycoproteins and O-linked glycosylation sites
PIR-ALN	Protein sequence alignments
PRINTS	Protein squence motifs and signatures
ProClass	Families defined by PROSITE patterns and PIR superfamilies
ProDom	Protein domain families
ProtoMap	Automated hierarchical classification of SWISS-PROT proteins
SBASE	Annotated protein domain sequences
SMART	Signalling domain sequences
SYSTERS	Protein clusters
Proteome Resources	
Aaindex	Physicochemical properties of peptides
REBASE	Restriction enzymes and associated methylases
SWISS-2DPAGE	2D-PAGE images and reference maps
Yeast Proteome Database (YPD)	Saccharomyces cerevisiae proteome

Retrieval Systems and Database Structure	
KEYnet	Keywords extracted from EMBL and GenBank
Virgil	Database interconnectivity
RNA Sequences	
5S Ribosomal RNA Databank	5S rRNA sequences
ACTIVITY	Functional DNA/RNA site sequences
Collection of mRNA-like non-coding RNAs	Non-protein-coding RNA transcripts
Database on the Structure of Large Subunit Ribosomal RNA	Alignment of large subunit ribosomal RNA sequences
Database on the Structure of Small Subunit Ribosomal RNA	Alignment of small subunit ribosomal RNA sequences
Guide RNA Database	Guide RNA sequences
Intronerator	RNA splicing and gene structure in Caenorhabditis elegans
Non-canonical Base Pair Database	RNA structures containing rare base pairs
PLMItRNA	Plant mitochondrial tRNAs and tRNA genes
Pseudobase	Information on RNA pseudoknots
Ribosomal Database Project (RDP)	rRNA sequences, alignments, and phylogenies
RNA Modification Database	Naturally modified nucleosides in RNA
SELEX_DB	Selected DNA/RNA functional site sequences
Small RNA Database	Direct sequencing of small RNA sequences
SRPDB	Signal recognition particle RNA, protein, and receptor sequences
TmRDB	tmRNA (10Sa RNA) sequences
tmRNA Website	tmRNA (10Sa RNA) sequences
tRNA Sequences	tRNA and tRNA gene sequences
UTRdb	5' and 3' UTRs of eukaryotic mRNAs
Viroid and Viroid-Like RNA Database	Viroid and viroid-like RNA and vHDV sequences
Yeast snoRNA Database	Yeast small nucleolar RNAs
Structure	
PDB	Structure data determined by X-ray crystallography and NMR
CATH	Hierarchical classification of protein domain structures
SCOP	Familial and structural protein relationships
ASTRAL	Analysis of protein structures and their sequences
BioImage	Searchable database of multi-dimensional biological images
BioMagResBank	NMR spectroscopic data from proteins, peptides and nucleic acids
CSD	Crystal structure information for organic and metal organic compounds.
Database of Macromolecular Movements	Descriptions of protein and macromolecular motions, including movies
Decoys 'R' Us	Computer-generated protein conformations based on sequence data
HIC-Up	Structures of small molecules ('hetero-compounds')
HSSP	Structural families and alignments; structurally-conserved regions and domain architecture
IMB Jena Image Library	Visualization and analysis of three-dimensional biopolymer structures
ISSD	Integrated sequence and structural information
LPFC	Library of protein family core structures
MMDB	All three-dimensional structures, linked to NCBI Entrez system
MODBASE	Comparative protein structure models
NDB	Nucleic acid-containing structures
PDB-REPRDB	Representative protein chains, based on PDB entries
PRESAGE	Protein structures with experimental and predictive annotations
Protein Motions Database	Motions of protein loops, domains and subunits
ProTherm	Thermodynamic data for wild-type and mutant proteins
RESID	Protein structure modifications

Transgenics	
Cre Transgenic Database	Cre transgenic mouse lines
Transgenic/Targeted Mutation Database	Information on transgenic animals and targeted mutations
Varied Biomedical Content	
CarbBank	Complex carbohydrate/polysaccharide sequences
Dbcat	Catalog of databases
DrugDB	Pharmacologically-active compounds; generic and trade names
HOX-PRO	Clustering of homeobox genes
LocusLink/RefSeq	Curated sequence and descriptive information about genetic loci
Molecular Probe Database	Synthetic oligonucleotides, probes and PCR primers
MPDB	Information on synthetic oligonucleotides
NCBI Taxonomy Browser	Names of all organisms that are represented in the genetic databases with at least one nucleotide or protein sequence
PubMed	MEDLINE and Pre-MEDLINE citations
Tree of Life	Information on phylogeny and biodiversity
Vectordb	Characterization and classification of nucleic acid vectors

INVESTMENT RESEARCH

HEALTHCARE

Biotechnology

Akhtar Samad, M.D., Ph.D.	(212) 514-2342
Jason Reed, Ph.D.	(212) 514-2341
Alan J. Tuchman, M.D.	(212) 514-2345
John Tonkin	(212) 514-2348

Medical Technology

Alan J. Tuchman, M.D.	(212) 514-2345

TECHNOLOGY

Telecommunications Equipment

Ayelet Oron	(212) 514-2305
Gilad Alper	(212) 514-2356

Diversified Technology

Rami Rosen	(972) 3519-9004

SPECIAL SITUATIONS

Alan J. Septimus	(212) 514-2317
Peter H. Vogel	(212) 514-2336

ASIA PACIFIC

Telecommunications and Special Situations

Sandia Shih	(212) 514-2358