

Preparing Your Life Sciences Organization for Bioinformatics

Perseid Software Limited
Needham, Massachusetts

July 2001

Table of Contents

1	EXECUTIVE SUMMARY	3
2	BUILDING THE “NEXT GENERATION” INFORMATION INFRASTRUCTURE	3
2.1	Return on Investment	3
2.1.1	Generating the Required ROI for the “New Biology”	4
3	ENTERPRISE INFORMATION SYSTEMS FOR BIOINFORMATICS.....	4
3.1	Complex Data Integration will be the Standard	4
3.1.1	Non-Stop Information Delivery and Management Requirements	5
3.2	Protecting Complex Databases.....	5
3.3	Next Generation Database Requirements.....	6
3.4	Massive Scalability and Reliability Requirements.....	6
3.4.1	Scalability Requirements	6
3.4.2	Reliability Requirements	7
4	INFORMATION SHARING ENABLES SPEED-TO-MARKET	7
5	ENTERPRISE ARCHITECTURE FOR BIOINFORMATICS R&D	7
5.1	Building Next Generation Infrastructure for the New Biology	8
5.1.1	Components for the Model Architecture.....	8
5.2	Information Technology for the New Biology	8
5.3	Model Enterprise Research & Development Environment	9
6	CONCLUSION	11
7	ABOUT PERSEID SOFTWARE.....	11

Figures

Figure 1	Data Architecture for the New Biology	6
Figure 2	A High-Availability Enterprise Bioinformatics Systems Architecture.....	10

1 Executive Summary

A new generation of Bio-information is emerging: Computational Biology, Genomics and Proteomics. To study the relationships of proteins to disease states requires a new generation of information systems and computational resources to manage extraordinary amounts of information. The next generation Bio-information revolution will demand a change in strategy for new drug determination and review. The complexity associated with discovering, testing and packaging new molecules requires information systems built on an infrastructure that must protect, manage, and share mission critical data.

Market leadership in the development and commercialization of drugs and therapeutics will evolve into the ability of a life sciences enterprise to store, process, distribute and manage complex bioscience, clinical trial and market research data. The ability to create, manage and maintain mission-critical information systems supporting these data will determine the next-generation pharmaceutical companies' speed-to-market and financial success.

This paper will recommend a "model" architecture for next generation Bioinformatics information management requirements including attributes for availability, security, data protection, data mobility, data repurposing, and data sharing.

2 Building the "Next Generation" Information Infrastructure

2.1 Return on Investment

Today, the applications that are expected to have the greatest impact on the productivity of research and development include proteomics, bioinformatics, predictive toxicology, and pharmacogenomics.¹ In the United States, companies spend an average of \$240 million on clinical trials required as a drug passes through the Food and Drug Administration (FDA) for final approval. Pharmaceutical companies have billions of dollars at risk if they cannot correctly match a pharmaceutical product to a target patient population.

Research activities of life science companies are focused on ascertaining new biological mechanisms for identifying or treating illness, injury and disease. As recently as 1995, only 500 receptors or enzymes in human cells were identified. By 2000, the number had increased to more than 10,000 with pharmaceutical companies investigating as many as 200 targets per year.

As a result, a new "focused" approach to pharmaceutical testing is emerging—the genetic approach. The genetic approach can be used to determine *a priori* patients who might be put at risk from a particular drug's side effects. Research by McKinsey & Company with pharmaceutical clients has documented the effects of lost earnings and potential impact of this new biology.^{1 2} A primary cause of lost earnings and revenue by pharmaceutical companies occurs when a very small segment of a target patient population experiences side effects from an otherwise effective pharmaceutical. Warner-Lambert, for instance, temporarily pulled Rezulin from the UK market and suspended its regulatory filing in Europe to investigate why the drug proved toxic to the livers of one diabetic patient in 60,000. Genetic diagnostics (pharmacogenetics) could increase the value of

¹ Manish Bhandari, Rajesh Garg, Robert Glassman, Philip C. Ma, and Rodney W. Zimmell, "[A genetic revolution in health care](#)," *The McKinsey Quarterly*, 1999, Number 4, pp. 58–67.

² Richard C. Edmunds III, Philip C. Ma, and Craig P. Tanio, "[Splicing a Cost Squeeze into the Genomics Revolution](#)" *The McKinsey Quarterly*, 2001, Number 2, Article at a glance

some pharmaceuticals by identifying which pharmaceutical class would be most effective for specific groups of patients.

2.1.1 Generating the Required ROI for the “New Biology”

To build the necessary Bioinformatics infrastructure for the new biology, the primary requirement is comprehensive data gathering and management beginning with basic research and extending to clinical trials and the approval stages of pharmaceutical development. The information technology solution that best meet the needs of emerging Bioinformatics research and development requirements is the enterprise integration of clinical, administrative and financial information which will create a complete picture of each patient and the performance of a pharmaceutical.

With hundreds of millions of dollars at risk during new drug development, meeting regulatory requirements in a timely way and speeding market launch are critical. Each day a new drug is on the market can add millions of dollars of revenue. These demands create time pressure on the entire value chain required to produce a new pharmaceutical. Data integration and coordination is required between internal research and development organizations, external firms that may have licensed the discovery, contract research organizations, regulatory bodies, (in multiple host countries) and target market distribution firms.

Since deploying the technology of the New Biology promises to transform the overall economics of developing and distributing pharmaceuticals, the question for life sciences companies isn't whether to invest, but how soon to invest and how much?

3 Enterprise Information Systems for Bioinformatics

3.1 Complex Data Integration will be the Standard

Within the research and development organization(s), each researcher needs immediate access to the latest research data and analytical results. Multiple clinical trials may need coordination at multiple research centers. These requirements generate a demand for complex information management tools and techniques that maximize the ability of research organizations to communicate and disseminate research results. The results must be communicated to multiple participants to meet the demanding and complex time schedules encountered in new drug research and development.

The solution to these complex requirements is a new information management architecture and delivery model. The architecture should:

- Integrate biological, genomic, computational, clinical, administrative and financial data into a comprehensive enterprise-scale information management solution
- Integrate all data developed during research and development, approval and marketing into a comprehensive enterprise-level “data warehouse”³
- Facilitate sharing of all key data by all researchers and research entities, not only the primary researcher and prime research organization
- Enable pervasive access to the R&D data warehouse
- Store all key documents involved in the research and development process
- Apply Good Manufacturing Process (GMP) standards of audit, control and verification to all versions of all data so that a comprehensive history of the research process emerges from the raw data

³ See “*Building Mission Critical Document Management Solutions for Global Pharmaceutical Companies*”, Perseid Software Limited, May, 2001, www.perseidsoftware.com

Information delivery architecture for the new biology should provide numerous benefits including:

- Ability to analyze biological, genetic and clinical trial data from a single repository
- Timely access to mission critical research, approval and marketing data
- Facilitate communication among researchers and marketing personnel during all aspects of discovery, testing and market introduction and after-market verification and validation of efficacy
- Improve analytical processes that result from having integrated clinical, financial and administrative data in a single repository during clinical trials
- Provide more comprehensive and complete documentation with full audit trails
- Enable more timely regulatory filings

3.1.1 Non-Stop Information Delivery and Management Requirements

Pervasive information management and delivery requires non-stop high availability computing and data management architecture. Given the complex research chain anticipated for the new biology, mission critical information systems supporting research and development efforts must be scalable, reliable and comprehensive in nature.

Scalable solutions are able to add computational, information delivery and data storage quickly and easily with no disruption to the research and development processes.

Reliable information delivery and management solutions minimize the downtime associated with backup and recovery of core databases within the R&D data warehouse and also provide continuous access to the central data warehouse through the Internet.

Comprehensive information delivery and management solutions can easily, rapidly and reliably interconnect multiple host computers into an integrated computational and data management “platform” supporting the new biology’s research and development demands.

This next generation computational environment will combine massive computational power with comprehensive analytical and reporting tools—all built upon a foundation of real-time, non-stop databases within the central R&D data warehouse.

3.2 Protecting Complex Databases

The new biology will demand types of data integration and complexity that is rarely found in pharmaceutical research and development today. Integration will be required for basic biological research data, images, analytical results, reports, clinical trial data, diagnostic and treatment data and disease management results from many patient populations. The integration of administrative, financial and clinical data with biological and genomic data will create a new state-of-the-art for data integration and information delivery within the pharmaceutical industry.

The key requirements are:

- Full integration of chemical, molecular, genomic, clinical, administrative and financial data on a new pharmaceutical product
- Pervasive and comprehensive reporting, *ad hoc* query and analysis tools creating an web-enabled Internet-based E-Infostructure®
- High availability database management systems to maintain the availability and integrity of the data warehouse

- Continuous availability of the data warehouse to each enterprise and party associated with the drug discovery and development processes

3.3 Next Generation Database Requirements

Pervasive and continuous enterprise access is a critical requirement for the model information management and delivery architecture. Data will arrive in different formats from different sources. The information delivery and reporting processes must present a uniform and easy-to-use application “user interface” to the enterprise data warehouse. “Metadata” or information about the contents of the R&D data warehouse must be continuously available and updated regularly so each data warehouse user is aware of the contents and state of data. Enterprise storage management systems must connect to all computing platforms within the central research and development facility so all data is available continuously with a high degree of integrity and reliability.

The central databases within the R&D data warehouse will contain tiers of information — from low-level raw transactions and biological “objects,” such as molecules to complex disease management summaries by patient populations. Figure 1 describes the tiers of information that can be expected to exist in central databases supporting the new biology.

A Data Architecture for the New Biology

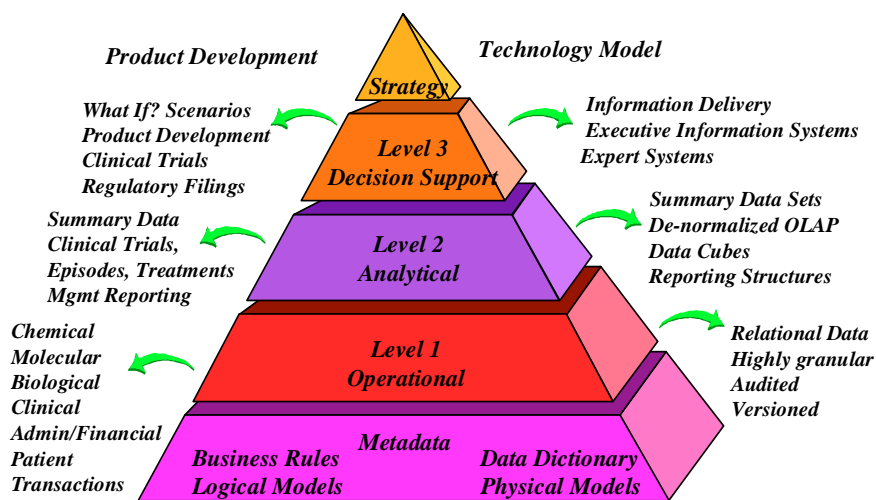


Figure 1 Data Architecture for the New Biology

3.4 Massive Scalability and Reliability Requirements

3.4.1 Scalability Requirements

Moving from chemical, molecular and protein data to an integrated database of clinical, administrative, financial and bio-information requires massive scalability. What may have worked on paper or in the internal research laboratory will no longer work in a global Bioinformatics research enterprise. The bottleneck in research is moving from simple computing cycles to integrated information and data management. Solutions

that worked with a few researchers and a small laboratory computer do not scale with terabytes of integrated bio-information.

The scalability requirements for the model architecture must include the ability to:

- Scale central processing platforms from departmental servers to massively parallel research computers
- Integrate multiple computing platforms with multiple operating systems
- Expand on-line real-time Internet-enabled reporting and “push” research information to researchers at multiple sites
- Grow enterprise databases across multiple computing platforms and operating systems

3.4.2 Reliability Requirements

Reliability requirements for the Bioinformatics enterprise research facility are substantial. Not only must individual hardware components be reliable, they must fully integrate into a uniform software environment that ensures continuous availability of enterprise Bioinformatics research solutions. “Downtime” is permissible, but only on the order of an hour or so per month.

The reliability requirements for the model architecture must include the ability to:

- Guarantee central processing availability with limited or zero down-time
- Integrate multiple computing platforms with multiple operating systems and easily upgrade network storage and central processors
- Ensure a high degree of integrity of Internet web sites for data entry, reporting and dissemination of research information
- Protect the integrity of enterprise database(s) across multiple computing platforms

4 Information Sharing Enables Speed-to-Market

Data will be aggregated by class and be continuously available for use in research, development and marketing. The opportunity for discovery of novel unintended uses for a pharmaceutical will require that all data be kept on-line and be immediately available.

The model architecture assumes that the following classes of data are kept in the central repository for ease-of-use and ease-of-access:

- Biological information such as chemical, molecular, genetic, and proteomic data
- Clinical trial data from all trials and all patients
- Regulatory documentation and research and development documentation
- Administrative, financial and clinical disease management and treatment data
- Insurance related claims and encounter data including all medical claims and services
- Medical diagnostic, treatment, and encounter data side-effect data

5 Enterprise Architecture for Bioinformatics R&D

The information technology required for the new biology is diverse with complex operational requirements. These requirements can be summarized as follows:

- Availability—Continuous access to the enterprise data warehouse and websites
- Security—Appropriate controls for access by role and certificate
- Data protection—Loss of data is unacceptable
- Data mobility—Data needs to migrate to the user who needs it, when it is needed and in the form required

- Data repurposing—Data and information have multiple purposes and roles
- Data sharing—Access to all information by all researchers and organizations, not just the primary reviewer

These requirements imply a level of reliability, availability, security, and access and control at or near the current state-of-the-art in computational biology and information management.

5.1 Building Next Generation Infrastructure for the New Biology

Very few information technology suppliers can currently meet the complex operational, reliability, availability and flexibility requirements anticipated for the information management demands of the New Biology. The most successful pharmaceutical companies will be those that invest in best of breed components — from each enterprise server, to enterprise storage systems, to database management systems and Bioinformatics applications that give the enterprise a competitive advantage. The bottleneck in moving from the current generation of information technology to the next generation in Bioinformatics is moving from simple “CPU power” to enabling the information and data management solutions for the new biology.

5.1.1 Components for the Model Architecture

The model information technology architecture for the next generation Bioinformatics research and development environment is composed of the following recommended classes of information technology:

- ❖ Computing Platforms and Operating Systems
 - Compaq Alpha
 - Sun Computers E-10000®
 - Unisys® E7000 Microsoft® Windows 2000 Data Center™
 - IBM S/390®
 - IBM AIX®
 - IBM MVS
 - IBM Z/OS
 - Microsoft Windows 2000
 - Sun Solaris®
 - Compaq Tru 64
- ❖ Storage Management Solutions
 - EMC® Symmetrix™
 - EMC Connectrix™
 - EMC SRDF™
 - EMC TimeFinder™
- ❖ Database Management Solutions
 - Oracle®
 - IBM DB2®
- ❖ Internet and Reporting Solutions
 - IBM Websphere®
 - SAS®
 - Microstrategies®
 - Cognos®

5.2 Information Technology for the New Biology

The goal of the information technology revolution for the new biology is pervasive and reliable access to research and development information whenever and wherever it is needed.

The foundation of the model architecture is based on EMC “E-Infostructure” composed of physical, connectivity and functional layers of hardware and software. The physical layer includes Symmetrix™ and CLARiiON™ enterprise storage management hardware that provides the basic foundation for performance, capacity, availability and other physical requirements of the central repository and database management systems. The Enterprise Storage Network (“ESN”) is the connectivity layer and through two information connections—Connectrix™ and Celerra™—it provides a means of using *all* primary and secondary operating systems to connect into the enterprise application and data management platforms. These could include each of the IBM operating systems, every major Unix operating system, including Linux and those from Compaq, Sun, HP, and IBM; and each Microsoft operating system.

5.3 Model Enterprise Research & Development Environment

The model architecture is based on a four-tier information technology architecture. Each tier can be replicated to enhance reliability and availability of the operational systems. The tiers of the proposed model architecture are:

- **Presentation** — Multiple devices with multiple modes of information delivery. Each device should be able to receive the mode(s) of information delivery suited to the device’s characteristics. For example, a PDA would receive formatted XML to map the display contents to the screen size of the PDA.
- **Network** — Wide-area and local-area redundant access to the computational and data management layers below and the presentation layers above. Wireless presentation access is included as a presumed capability of the network layer.
- **Computational** — Computational, analytical, application-oriented servers providing the services needed for the model architecture on a “best-of-breed” basis. Highly scalable central processors with high-availability operating systems, e.g., IBM Z/OS and AIX HACMP. Inexpensive processors for secondary applications, e.g., Intel-based processors for Windows 2000 applications.
- **Data Base Management**—Oracle Parallel Server or IBM DB2 to support the databases for the central repository
- **Enterprise Storage Management** — The storage management layer composed of the Enterprise Storage Network (“ESN”) and the Database Management System(s). EMC Symmetrix™ and enterprise storage management software to integrate all operating systems and database storage into a uniform central repository. EMC Connectrix storage management to handle multiple connections to the servers. EMC TimeFinder™ can provide each researcher with his own copy of data to analyze, thereby increasing productivity. TimeFinder can be used to refresh data warehouses with timely information without disrupting production systems. The remote mirroring capabilities of SRDF™ can protect clinical trial and other critical data to avoid costly interruptions in speed to market.
 - EMC TimeFinder for non disruptive backup and data warehouse loading
 - EMC SRDF for disaster recovery and information mobility

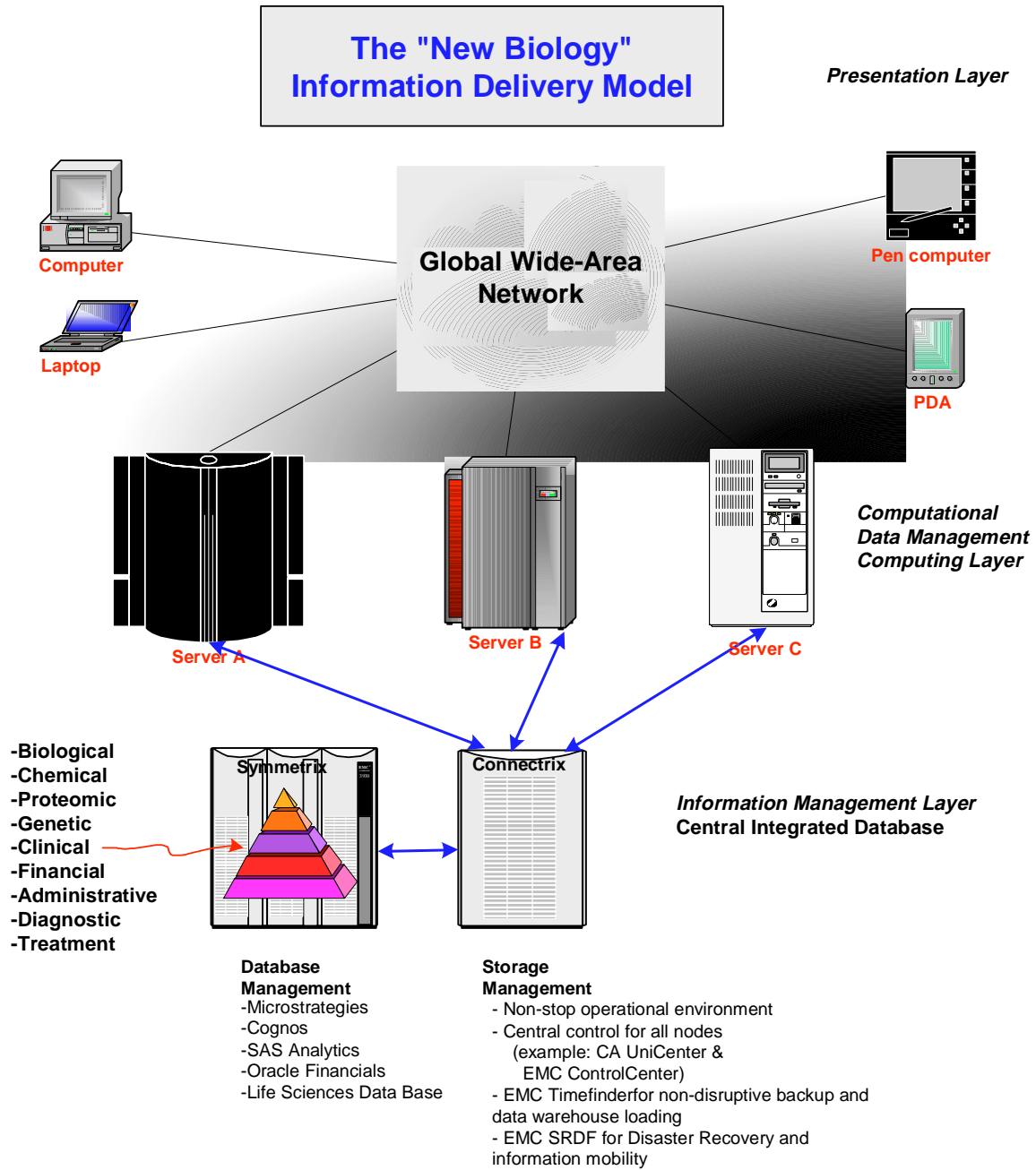


Figure 2 A High-Availability Enterprise Bioinformatics Systems Architecture

6 Conclusion

The New Biology is about rapid discovery and deployment of pharmaceuticals into a complex and growing global market—at great risk and expense. Computational Biology, Genomics and Proteomics will drive developments in information systems technology for emerging Bioinformatics applications.

This paper recommended a “model” architecture for the “next generation” Bioinformatics information management requirements that included attributes of availability, security, data protection, data mobility, data repurposing, and data sharing. These attributes form the foundation of next generation Bioinformatics solutions because, in combination, they ensure trust and encourage reliance on the core information technology for research and development.

The model delivery architecture ensures that evolving information about a pharmaceutical in the market is both continuously available and accurate—key goals and objectives of the New Biology.

From a business perspective, the next generation Bioinformatics solutions will focus on integration. Integration of key administrative, clinical and financial data will ensure that the entire life-cycle of a pharmaceutical is available to all researchers and marketers at all times.

Finally, the speed-to-market for each product developed from the New Biology is critical. Business and information technology requirements must be well articulated and their solutions executed with precision if the Bioinformatics revolution driven by the New Biology is to generate immediate and persistent returns-on-investment for the aggressive New Biology adherents.

Execution will be a challenge, but those who possess both vision and ability to execute will overtake the timid.

7 About Perseid Software

Perseid Software is engaged in providing strategic consulting and information technology design services to healthcare and life sciences enterprises. For more than 30 years, the principals of Perseid Software have been engaged in the development of mission-critical information systems and in the analysis of healthcare, disability and pharmaceutical data.

Perseid Software is not merely a strategic consulting firm. It is an engineering management and design firm focusing on database design and implementation of very large and complex life sciences and healthcare information systems. Perseid’s clients include or have included some of the largest and most progressive computer, healthcare and manufacturing companies in the world.

Contact:

Bernard P. Wess, Jr., President
Perseid Software Limited
Needham, MA
Direct Dial:(781)453-2351
bwess@perseidsoftware.com
www.perseidsoftware.com

IBM®, z/OS™, OS/390™, AIX™, MVS™, AS/400®, AIX/HACMP™ are trademarks of the IBM Corporation
Sun Solaris™ is a trademark of Sun Computers, Microsoft®, Windows NT™, Windows 2000/Data Center™ are trademarks of Microsoft Corporation, Symmetrix™, CLARiiON™, Connectrix™, Celerra™, Symmetrix Remote Data Facility™, TimeFinder™ Software, EMC Foundation Suite™ and Database Edition for Oracle™ are trademarks of EMC Corporation, E-Infostructure® is a registered trademark of EMC Corporation
Oracle Open Parallel Server™ is a trademark of Oracle Corporation, Compaq® is a trademark of Compaq Computers, Unisys® is a trademark of Unisys Corporation