The promise of a virtual lab in drug discovery

Han Rauwerda[1], Marco Roos[2], Bob O. Hertzberger[3] and Timo M. Breit[4]

[1]Integrative Bioinformatics Unit, Institute for Informatics, Faculty of Science, University of Amsterdam, Kruislaan 318, building 1, room C017, P.O. Box 96062, 1090 GB Amsterdam, The Netherlands
[2]Integrative Bioinformatics Unit, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Kruislaan 318, building 1, room C017, P.O. Box 96062, 1090 GB Amsterdam, The Netherlands
[3]Computer Architecture and Parallel Systems Group, The Computing, System architecture and Programming Laboratory, Institute for Informatics, Faculty of Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
[4]Integrative Bioinformatics Unit & Micro-Array Department, Swammerdam Institute for Life Sciences, Faculty of Science, University of Amsterdam, Kruislaan 318, building 1, room B120, P.O. Box 96062, 1090 GB Amsterdam, The Netherlands

To date, the life sciences 'omics' revolution has not lived up to the expectation of boosting the drug discovery process. The major obstacle is dealing with the volume and diversity of data generated. An enhanced-science (e-science) approach based on remote collaboration, reuse of data and methods, and supported by a virtual laboratory (VL) environment promises to get the drug discovery process afloat. The creation, use and preservation of information in formalized knowledge spaces is essential to the e-science approach. VLs include Grid computation and data communication as well as generic and domain-specific tools and methods for information management, knowledge extraction and data analysis. Problem-solving environments (PSEs) are the domain-specific experimental environments of VLs. Thus, VL–PSEs can support virtual organizations, based on the changing partnerships characteristic of successful drug discovery enterprises.

'Omics' technologies have changed the arena of life sciences research forever. They allow generation of data at a large scale, starting with whole-genome sequencing and followed by microarray gene-expression analysis and mass spectrometry of proteins and metabolites. Data are produced at a startling rate by a constantly growing number of new high-throughput and/or genome-wide biotechniques at each cellular level: genomics (DNA), transcriptomics (RNA), proteomics (protein), metabolomics (metabolite) and phenomics (phenotype). With nanotechnology and laboratory-on-a-chip applications the end of this development is not yet in sight.

Omics technologies require substantial paradigm shifts for the way life sciences research is carried out (Table 1) [1]. For instance, instead of gene-by-gene analysis, whole genomes can be analyzed. This steers life sciences towards a more holistic, or 'systems', as well as data driven approach. Omics experiments are relatively expensive, which forces scientists to focus on advanced design for experimentation as part of a whole-chain research approach. Furthermore, the data generally contains information outside the scope of the original experiment. Hence, to maximize the proceeds, omics data needs to be reusable, shareable and suitable for *in silico* experiments. All of this poses high demands on annotation of data and standardization of data formats. Furthermore, the conversion into information and knowledge to support scientists answering biological questions requires advanced analysis methods and tools that enable mining and integration of these complex datasets [2].

The bottlenecks for life sciences have shifted from data generation to data storage, preprocessing, analysis and interpretation. The challenge is to remove these bottlenecks by a combination of life sciences and information technology (IT). The area that deals with this challenge is called enhanced-science (e-science) and is characterized by multidisciplinary collaborations [3,4]. Some key aspects here are remote collaboration, data and resource sharing [5], data integration, information management and knowledge handling. The integration of methodologies and infrastructure needed

Corresponding author: Breit, T.M. (breit@science.uva.nl).

1359-6446/06/$ – see front matter PII: S1359-6446(05)03696-2

for e-science experimentation form the basis for the concept of a virtual laboratory (VL) [6].

This review will describe the origin, concept, structure and promise of a VL in the broader context of life sciences and informatics R&D. We will also indicate the specific relevance and promise that VLs could hold for drug discovery.

Virtual laboratory drivers

Biological, economical, and information and communication technology drivers

In contrast to their longstanding tradition in the development of biotechnology, life scientists have little experience in dealing with large information-rich datasets. Yet, especially when integrating these datasets, a system can be studied as a whole (systems approach), patterns can be recognized, models tested, networks of relationships built, biomarkers for diagnostics identified and leads for drug development discovered. Ideally, life scientists will work in an e-science environment in which they can conveniently perform these tasks to design experiments, build models, indicate the tenability of a hypothesis, or be notified in case of availability of distant data or methods [7].

The increasing complexity of science makes it economically impossible to build up all the necessary skills and competence at one location [4], which is also true for research laboratories of multinational companies that operate on a global scale. In general, e-science environments should provide an IT infrastructure that allows for reuse and sharing of data, methods, experimental designs and knowledge within a setup of (*ad hoc*) collaborations, in which the whereabouts of scientists and data are not important.

All of this is feasible due to the recent development of high-performance networking [8] using optical network technology that makes resources at remote locations accessible. High-performance networking thus enables new, communication-intensive ways of working (e.g. pooling computational resources, streaming large amounts of data from databases or instruments to remote computers, linking sensors to each other and to computers and archives, and connecting resources in collaborative environments). At the same time it will allow high-performance and high-throughput computing with so-called parallel computers in a distributed environment like the computational Grid.

Drug discovery driver

The omics revolution has already had a major impact in drug discovery with respect to target identification, selection and validation because a better understanding of phenomena at the molecular level can improve the whole drug discovery process [9–14]. It holds the promise of larger numbers of more efficacious drugs with fewer adverse effects at a lower total cost [12]. Despite this, the drug developing industry has been struggling with overall low productivity in recent years. Research costs are exploding and the number of new drug applications is falling [15]. There is little debate about the usefulness of omics technology in drug discovery [16–18], even though the ability to generate data has outpaced the ability to understand it, which limits its current applicability to drug discovery. Even more than the life sciences community, the drug discovery community is in the relatively slow process of integrating omics technologies into research [13] because huge investment is required to adjust to the omics paradigm shifts (Table 1), and time is a coercive factor. This is compounded by the fact that drug discovery deals with the complexity of a whole chain – from the molecular to the organism level – simultaneously. Nevertheless, industrial drug discovery data are usually of high quality, because of ample resources and laboratory standardization, good regulation, a well-defined pipeline and a distinct final goal, all of which form an extremely good basis for application of an omics approach in drug discovery. Recent literature about omics' usefulness in drug discovery provides a wide range of potential solutions, from outsourcing to adopting a systems or integrative biology approach [19–26]. However, although none of these solutions will provide the ultimate answer to today's drug discovery problems [27], they all share elements of collaborated use of distributed resources, reuse of data and methods and computer support. This is exactly what VLs are about.

TABLE 1

Trends in life sciences research due to the omics revolution.

	Pre-omics era	Omics era
Data	Limited parameters	Genome-wide
	Small datasets	Massive datasets
	Limited data reuse	Extensive data reuse
	Limited data sharing	Extensive data sharing
	Poorly annotated	Use of metadata and standards
Experimentation	Relative inexpensive	Expensive
	Hypothesis-driven	Data-driven
	Simple design	Advanced design
	Wet-laboratory	*In silico*
Analysis and interpretation	Straight-forward	Complicated
	Basically manual	Strongly computer aided
Research approach	Segmented	Whole-chain
	Reductionistic	Holistic
	Mono-disciplinary	Multi-disciplinary
Science	Experimental	Conceptual
	Limited collaborations	Extensive collaborations
	Long-standing collaborations	*Ad hoc* flexible virtual organizations
	Conventional	e-Science

Why is it taking so long to develop a virtual laboratory?

Despite convincing pull-and-push factors, it is taking a long time to develop a VL. Much isolated work is being done on e-science applications and networking infrastructure. However, to obtain the envisioned e-science environment, the entire chain – from applications via generic software components to networking – must be studied and built collaboratively and systematically. Then again, building and implementing is generally not considered a task for academic research groups and, moreover, it takes time to bridge the gap between scientists from different backgrounds who often are subject to the conflict between common goals of the collaboration and short-term goals demanded by their own field. This catch-22 situation slows down progress but the potential benefits will eventually outweigh current obstacles. Funding of e-science projects as such partly takes away such obstacles.

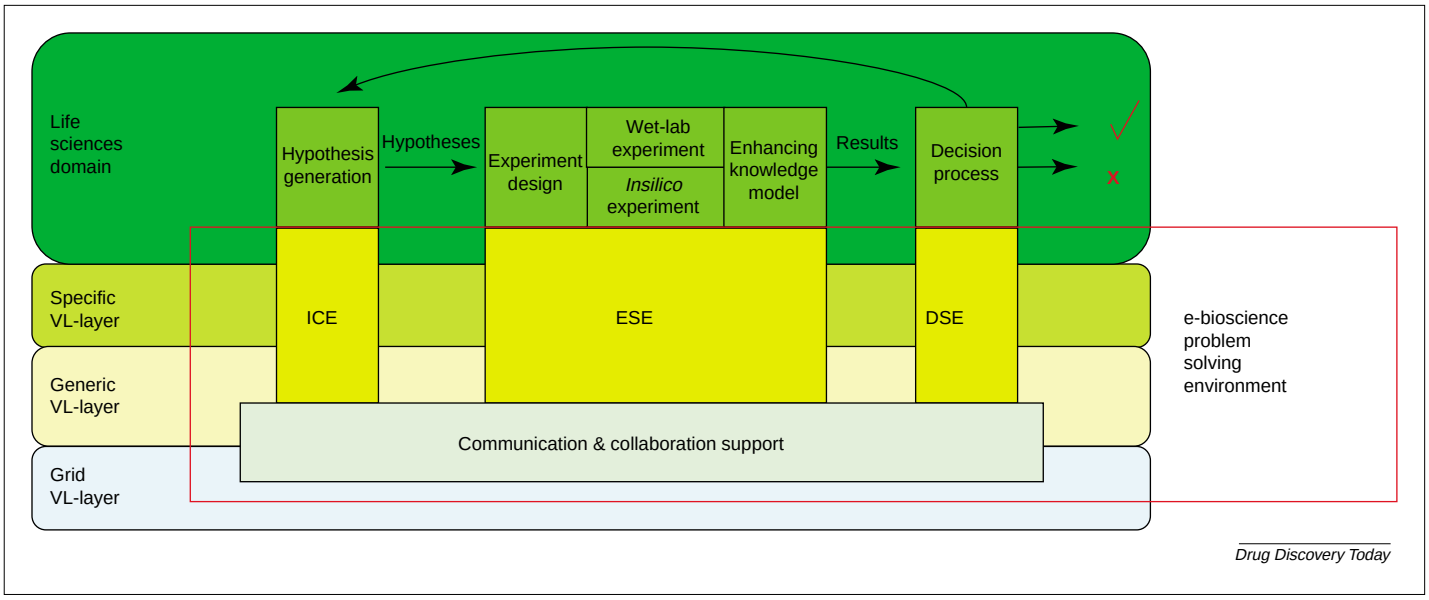

FIGURE 1

A concept of a problem solving environment for e-science experimentation in life sciences (e-bioPSE). A virtual laboratory (VL) provides tools, methods and infrastructure to create an e-bioPSE. A VL consists of three stacked layers: a Grid VL layer for large scale distributed computing; a generic VL layer that provides the generic methods, tools and infrastructure; and a specific VL layer that interfaces between the application domain and the generic VL layer. An e-bioPSE stretches from the specific application domain into the Grid VL layer and comprises several supporting environments (see main text for details) defined by the demands of the entire application research chain as indicated. These PSE elements are assisted by communication and collaboration support. Abbreviations: ICE, interactive and creative environment; ESE, experimentation support environment; DSE, decision support environment.

Conceptual framework

In the conceptual framework, e-science is about the collaborative and computer-aided approach of any science. Thus, e-science applications are the drivers for, and the users of, VLs. VLs provide the tools, methods and infrastructure to enable e-science experimentation. Problem-solving environments (PSEs) are the domain-specific experimental environments equipped with generic and specific VL parts and populated by virtual organizations (VOs) (Figure 1).

General concept of e-science

The term 'e-science' was introduced around 1999 by John Taylor: 'e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it' [4].

E-science can signify digitally enhanced, or electronic, science, originally defined by being at least based on high-power computing and high-speed networking [4,28]. It is a paradigm for doing science in global collaboration enhanced by advanced computing and communication technologies and driven by floods of data coming from a variety of sources.

The heart of any science domain is its knowledge space (Figure 2), which has been defined as 'the sum of all types of (proprietary) data and information within the scope of interest and composed of relevant databases, information sources, document/knowledge bases, metadata and a knowledge map' [29].

The success of an e-science approach depends on the support for such a knowledge space. Collaborative use of it by humans and computers must be accommodated. Therefore the knowledge space must be formalized, that is, based on a shared and computer-readable conceptualization of a domain. Hence e-science collaborations will stimulate the creation and preservation of formalized knowledge spaces by the use of ontologies and through the implementation of federation schemas or semantic webs [30]. Ontologies are a formal way of representing knowledge in which concepts are described by their meaning and their relationship to each other [31]. Formalized knowledge spaces allow for reuse of elements from other formalized knowledge spaces, which is even more important because science is increasingly carried out in globally distributed and frequently changing collaborations (i.e. virtual organizations, as detailed later).

General concept of a virtual laboratory

At the *Expert Meeting on Virtual Laboratories* (Ames, IA, USA, 1999) organized by the International Institute of Theoretics and Applied Physics, a VL was broadly defined as 'an electronic workspace for distance collaboration and experimentation in research or other creative activity, to generate and deliver results using distributed information and communication technologies'.

Furthermore, VLs were considered to be neither replacements nor competitors for real laboratories, but rather as extensions that hold new opportunities not realizable within a real laboratory at an affordable cost. Alternative terms associated with the VL concept include 'collaboratory' and 'distance collaboration group'. William Wolfe, who coined the word collaboratory in 1989, defined it as 'a center without walls, in which the nation's researchers can perform their research without regard to geographical location – interacting with colleagues, accessing instrumentation, sharing data and computational resources and accessing information in digital libraries'.

Since then, the VL concept has been expanded to advanced opportunities for integrated teaching, research and promoting cross-disciplinary research. Essentially, VLs are about moving towards dynamic and *ad hoc* organizational structures in society and as such they can encompass almost all spheres of human intellectual endeavor.

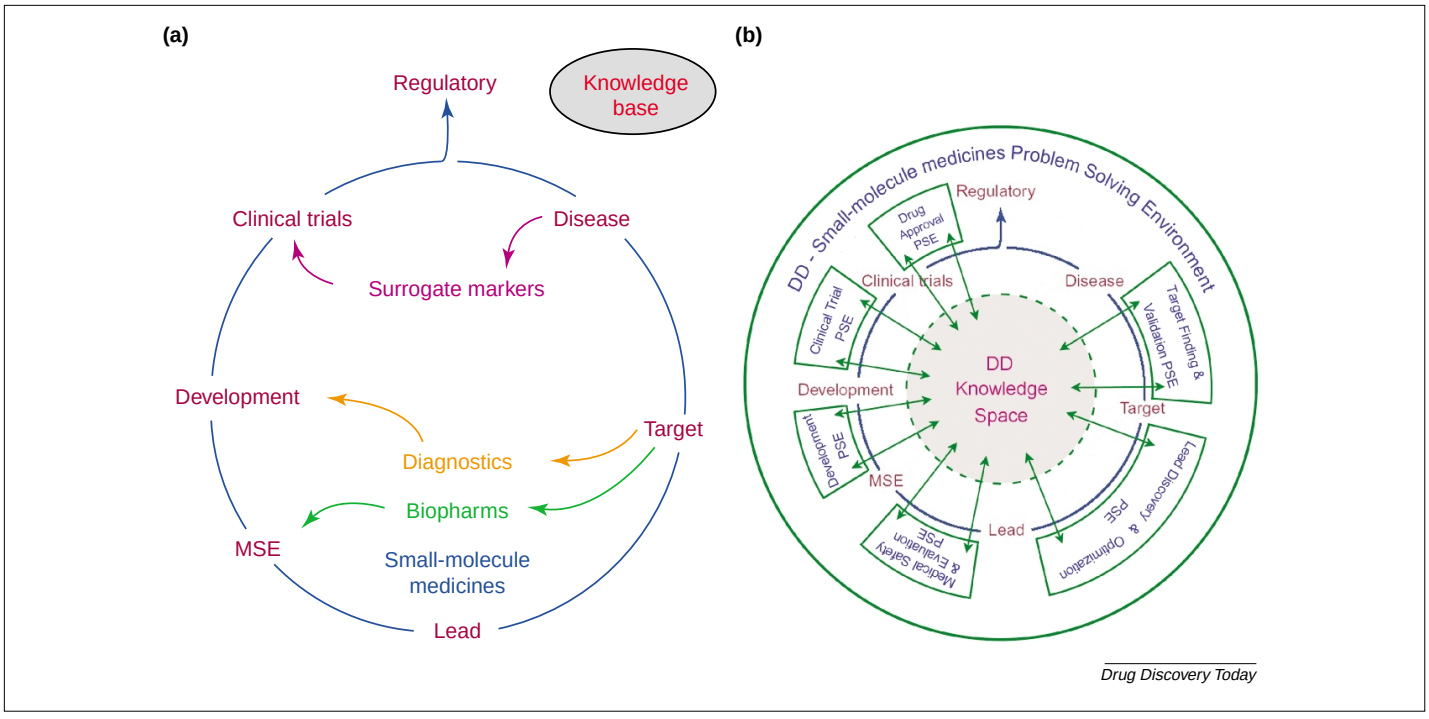

FIGURE 2

A problem-solving environment in drug discovery. (a) The information flow of five pharmaceutical product classes: small-molecule medicines, biopharmaceuticals, diagnostics, surrogate markers and knowledge bases. The path for small-molecule medicine is shown in black, whereas the other paths are shown in the color of their product class. **(b)** Each phase in the drug discovery cycle for small-molecule medicine is supported by a dedicated sub-PSE in the context of a master PSE. The sub-PSEs use and augment one knowledge space that can also be used by PSEs for other pharmaceutical product classes (not shown). Generic VL methods of any sub-PSE can be reused in any other sub-PSE. Abbreviation: MSE, medical safety and evaluation. Adapted, with permission, from Ref. [43].

The fundamental principles of VLs are collaboration together with data and resource sharing in a manner independent of time and place [32]. In life sciences, VL development is boosted by the new omics research adage – 'collaborate or perish'. Features of a VL comprise the following: they are based on the exchange of methodologies from different science areas and, as such, are highly multidisciplinary; they can house virtual organizations of public and private partnerships; they demand controlled data and resource sharing in a secure environment to promote reuse of valuable data; they cover the complete e-science technology chain from applications to networking; they comprise a distributed environment, as required by the time and place independency; and they aim for generic solutions that result in reusable components. Once the basis of a VL is established, it will be an ever-expanding entity by the introduction of new applications, additional methodologies and various collaborations.

General concept of a problem-solving environment

The term PSE originally emerged in the 1960s with the first high-level computer languages, but the limited computer power then prevented any practical use. PSEs are commonly described as computer systems that provide all the computational and methodological facilities necessary to solve complex problems in a specific domain [33]. A characteristic of PSEs is that they use the language of the problem domain, so they can be used without specialized knowledge of the underlying computer hardware or software [34,35]. Straightforward PSEs are software solutions to a specific area, such as SAS® and SPSS® for statistics. In the context of a VL, a PSE acquires enormous potential because it harnesses the power of the underlying VL elements. For example, achievements in distributed Grid computing can be used in a secure environment that is dedicated to a specific virtual organization, such as a drug discovery partnership. Problem solving and experimentation are enhanced by: (i) workflow tools for designing experiments; (ii) environments in which data can be handled securely; (iii) tools to integrate data from different sources; (iv) access to data and information by data mining and visualization methods; (v) knowledge handling facilities (e.g. for modeling); and (vi) facilities, such as data sharing and notification, for collaboration in virtual organizations. PSEs are domain-specific and expandable; they are built by 'e-science aware' domain experts, middleware experts and Grid experts by reusing generic technology, developing missing VL parts and creating domain-specific components.

General concept of a virtual organization

VOs are dynamic and *ad hoc* organizational structures made up of joint ventures, industrial partnerships and subcontractual arrangements. Hence, a VO is any pattern of organization based around distributed physical, human and knowledge resources, and – most usually – tied together with information technology systems that enable such resources to perform value-added activities [36]. This enables disparate groups of organizations and/or individuals to share resources in a controlled fashion, so that VO members can collaborate to achieve a shared goal [37]. This started nearly two decades ago in high-energy physics with the emergence of 'big science'. Now these flexible collaborative networks are also emerging in life sciences.

TABLE 2

Important methodologies for each layer of a virtual laboratory

VL layer	Methodology	Concerns
Specific	End-user interface	Providing easy access for novice users to PSE
	Text mining	Retrieving information from literature and other text sources
	Data mining	Retrieving information from data sources
	Data analysis	Turning data into information and knowledge
	Data integration	Associating individual data sources
	Semantic modeling	Modeling the domain knowledge
Generic	Visualization	Displaying information in a user configurable way
	Workflows	Methods and infrastructure to support for process flows
	Information disclosure	Dynamic model driven information and knowledge extraction
	Information management	Data and meta-data federation to store and manipulate data
Grid	AAA	Authentication, authorization, and accounting
	Data storage	Distributed data storage
	Storage brokerage	Mediating access to diverse data sources
	Cycle scavenging	Computing by using idle resources
	ICT infrastructure	Networked, large distributed environment

Virtual laboratory structure

A VL can be divided into three stacked layers: a specific VL layer, a generic VL layer and a Grid VL layer. The specific VL layer interacts with the individual application domains (Figure 1). The generic VL layer enables the specific VL layer to use complicated Grid and network technology of the Grid VL layer. Each layer has its own methodologies (for example, see Table 2), as well as its accompanying information and communication technology (ICT) infrastructure to support specific demands. Although boundaries between these layers are fuzzy, as a rule, methodologies and ICT infrastructure become more generic towards the lower layers. Hence, increasing reuse of methodology at lower layers by higher layers is an important feature of a VL. In essence, a VL is a collection of tools and methods which are used at choice in any given PSE that constitutes the experimental environment for a specific domain.

The Grid virtual laboratory layer

Large-scale distributed computer environments involve the actual networking services and connections of a potentially unlimited number of ubiquitous computing devices. Such a distributed infrastructure is, in analogy to the electric power grid, often referred to as a 'Grid' [4]. Ian Foster and Carl Kesselman, pioneers of the Grid, define it as an enabler for virtual organizations: 'an infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources' [37,38].

Grids are often classified by their content and level of integration as: (i) computational Grid; (ii) data Grid; (iii) information web or Grid; and (iv) semantic web or Grid [3,39–42]. A major purpose of the technology in the Grid layer of a VL is to manage the resources necessary for e-science. It shields computer operating specifics from the user, allows resource management and facilitates access to shared resources between scientists in a secure fashion. The necessary Grid infrastructure will inevitably be substantially more complex than the infrastructure used to serve pages on the Web. Resources here include not only entities like computing power, memory capacity or communication speed, but also abstractions such as data storage, information sources and physical resources like mass spectrometers.

The generic virtual laboratory layer

The generic VL layer is at the heart of any VL. It provides the generic methods, tools and infrastructure to use Grid technology to meet the demands from the applications [6]. Because many different applications feed into the generic VL layer by using their own PSE, the available methods and techniques must be generically applicable. Elements of this layer include the generic parts of: (i) architecture to support interactive high-performance and high-throughput computing; (ii) collaborative information management methods, such as metadata catalogues and ontologies; (iii) workflow processing tools; (iv) adaptive information disclosure by a suite of dynamic model-driven information and knowledge-extraction methods and techniques; (v) advanced visualization tools; and (vi) intuitive end-user interfacing. The methodology and functionality of the generic VL layer forms the basis for the application-specific PSE. In essence, this layer should hold all the necessary tools and expertise to build a PSE. Tools must be as generic as possible to allow for easy plug-in capability.

The specific virtual laboratory layer

The specific VL layer provides the interaction between the application domain and the generic VL layer. In a way, the specific VL layer will allow the application domains to adopt an e-science approach. This is done by creating domain-specific PSEs. This layer consists purely of the domain-specific parts of PSEs. It provides domain-specific methods, tools and infrastructure, which are either adapted from the generic VL layer or are created. Elements of this layer are the specific parts of those mentioned in the generic VL layer.

Problem-solving environment elements

A PSE stretches from the specific application domain into the Grid VL layer (Figure 1) and comprises several supporting environments

defined by the demands of the entire application research chain. Different PSEs relate on a technical level through generic elements from the generic VL layer and the Grid VL layer and communicate on a content level via the formalized knowledge space (Figure 2). The components of a PSE supporting four different phases of a life sciences research cycle that is applicable to drug discovery are discussed below (Figure 1).

An interactive and creative environment

A whole-chain research process starts with an exploration of the problem domain. The purpose is to make an inventory of relevant data, information, knowledge and ideas often hidden in databases, literature and the minds of experts, and then use this in the creative process of defining the problem, generating new hypotheses and drawing-up primary models. These processes are carried out in face-to-face, multidisciplinary collaboration between e-scientists and problem owners, such as molecular biologists and medical doctors. The interactive and creative environment (ICE) supports all of this with: (i) information retrieval methods to disclose information; (ii) semantic web technology to provide a common framework for referencing data sources and their associated metadata; (iii) bioinformatics tools to perform explorative analyses and visualize information; (iv) knowledge-capture methods to model information; and (v) grid technology to support collaboration between scientists from different disciplines and at remote locations.

The experimentation support environment

In the next e-science phase, the actual wet-laboratory and *in silico* experiments are designed, executed and interpreted in the context of modeled domain knowledge. This involves the generation of data and related metadata, data modeling, computational experimentation and associated methods. This again requires an interactive environment for collaborating with multidisciplinary teams of experts. The experimentation support environment (ESE) provides all necessary data integration tools, metadata-related services and content-driven modeling methods for actual experimentation. It will do so in a highly configurable way, allowing researchers to embed the process from design for experimentation (up to the enhancement of knowledge models) in workflows. Information about version, date, location and type of the data and services used (i.e. provenance data), will be captured from the workflow tool. Additionally, notification services can be set up to inform project teams when services or data are modified or when new results are available [43]. To be able to make these large amounts of heterogeneous data in the PSE interoperable, data needs to be annotated, for example, using ontologies and Life Sciences Identifiers [7,44].

The decision support environment

The decision phase – whether to enter the next stage in a research project or to phase out – is carried out at a managerial level. To make the right decisions, it is crucial to have the right information at the right time [45]. The decision process can benefit from a wide range of support technologies, such as report generation tools, search interfaces, alerts generators, or rule-based or machine learning engines [46]. Because decision takers are often pressed for time and lack the expertise to work with complex scientific tools, the decision support tools should have a high level of automation and provide comprehensive summaries with convincing facts. These facts come from within and from outside the PSE. The decision support environment (DSE) is able to deal with all these requirements and is still accessible through a user-friendly interface.

Communication and collaboration support

Communication and collaboration (CandC) are essential in the functioning of VLs. Types of CandC in a VL and means by which they are enacted are: (i) CandC between experts by enabling easy contact plus data and information exchange [47]; (ii) metadata communication by standardization and ontology-based modeling [7,31]; (iii) communication between data and experimenter using a wide array of visualization tools; (iv) communication between computer tools and/or environment and experimenter by means of intuitive end-user interfacing tuned to the level of the user; and (v) communication between computer resources via Grid technology. All components of a PSE must deal with these types of CandC to varying degrees. Because the support elements of CandC are generic parts of the VL, they are tackled in the generic VL layer and the Grid VL layer.

Virtual laboratory-like examples

Elements of the basic VL concept have successfully been implemented in some areas in life sciences. Some VL-like examples are:

- The cancer biomedical informatics Grid (caBIG) [48–50] is a grid environment primarily aimed at cancer research. The core of caBIG is a set of biomedical object models, implemented in Java middleware, which is connected to the US National Cancer Institute (NCI) databases and realizes data federation.
- The biomedical informatics research network (BIRN) is about data federation with a focus on neuroimaging [51]. Data from 26 research sites is managed through the BIRN virtual data Grid and is federated by a mapping to shared knowledge sources using ontologies and spatial atlases. Tools for analysis and visualization are made available through a single sign-on Java Grid interface.
- The myGrid project is a dedicated and independent PSE and is an open-source service-oriented environment for bioinformatics [39]. It supports virtual organizations and the use of distributed resources. MyGrid offers the middleware to perform *in silico* experiments, which consist of distributed queries executed through internet services, and workflows [43,52].
- The virtual laboratory for e-science (VL-e) develops e-science PSEs for several scientific areas, such as high-energy physics, medical imaging and bioinformatics. Thus, VL-e is about high performance computing, real-time processing, data federation, workflows and semantic modeling [46]. By developing middleware that is as generic as possible, elements and data will be reusable.

Drug discovery problem solving environment

A VL can also be employed by the drug discovery domain by using a drug-discovery-specific PSE. The life sciences process flow as depicted in Figure 1 also applies to each individual part of the cyclic drug discovery process (i.e. target identification and validation, lead discovery and optimization, medical safety and evaluation, drug development, clinical trials and drug approval) (Figure 2a) [45]. At the same time, various parts are used by pharmaceutical

product classes (i.e. small molecule medicine, biopharmaceuticals, diagnostics and surrogate markers) [45]. Thus, there will be numerous dedicated drug discovery sub-PSEs (Figure 2b). Although this might appear overwhelming at first, it is this modularity that makes the concept of a drug discovery PSE feasible [6]. When the basic framework of a VL and a drug discovery PSE is established, the individual drug discovery sub-PSEs can be created separately. Following the VL concept, the generic methods of any sub-PSE can be used in any other sub-PSE. This will actually speed up the development of a drug discovery PSE over time and, as such, it is an ever-expanding resource for its users. It will promote data and information reuse and will preserve and augment the capital of drug discovery, that is, the drug discovery knowledge space. This becomes increasingly important because drug discovery virtual organizations operate on a global scale and change frequently due to mergers, acquisitions and outsourcing [21,22].

Because the VL approach is a scalable concept, every (virtual) organization can in principle develop their drug discovery PSE. However, for the time being, building a functional VL environment is such an endeavor that it only can be developed in a large-scale collaborative manner. Drug discovery PSEs are about communication and collaboration along the entire chain of the drug discovery process and about creating a durable knowledge space. In a sense, analogous to the underlying utility Grid concept, PSEs allow access to data, information, and knowledge Grids, plus the appliances and computational power to use them.

Although no drug discovery PSEs currently exist, and the requirements for such PSEs will have to be worked out, it is clear that drug discovery PSEs will share facilities for high performance computing, tooling, data-plus-workflow management and collaboration. Some examples are presented below to illustrate the benefits of a VL at two stages in the drug discovery cycle and for the development of a drug discovery knowledge space.

Target finding and validation

One reason for the recent decrease in the number of new molecular entities (NMEs) [15] could be that targets, identified by animal models and genomic analyses, are not relevant to the studied disease [11]. An approach to overcome this is presented by the 'High-Throughput Human Disease-Specific Target Program' [11], which uses a single algorithm to combine single nucleotide polymorphism genotyping of patients with their complete clinical and demographic information. Embedding and using gene–disease association methodologies in a drug discovery PSEs will have the advantage that the obtained data, information and knowledge can easily be used in conjunction with analysis methods already available in VL. In the ICE, this involves knowledge representation and visualization [20], which enables introduction of external data, literature and clinical models of the disease for simple correlation experiments and multidisciplinary discussions. This will generate insight into the mechanisms of action of a disease that can corroborate target identification. In the ESE it will be feasible to easily evaluate and use, for example, several state-of-the-art statistical algorithms for genetics. In the DSE, the entire target finding route can be overseen because VLs are workflow-based and able to hold provenance data. Thus the decision process can be based on scientific information and knowledge. The gap that exists between superficial 'Powerpoint bullet'-style feedback and detailed monthly reports [53] is thus effectively closed and drug discovery teams can be adequately informed [43].

Lead discovery and optimization

The question Hodgman [45] raises – 'Why screen compounds that others have found to be toxic? – reflects current information management problems in lead discovery and optimization. HTS has moved from trying to optimize lead discovery by using ever-expanding compound libraries, towards high quality libraries that consist of (pure) chemicals, permutated around several promising scaffolds, pre-selection of compounds [16] and targeted libraries [54]. Optimizing this process requires introduction of information about structure and function of proteins and ligands. This, in turn, is only feasible with comprehensive information and knowledge management tools, such as ontologies, which will be available in a drug discovery PSE. Another obstacle in lead discovery and optimization is the demand for computing power for structural-based virtual screening methods [55–57]. The applied docking algorithms are computationally demanding, although the calculations can be distributed over many computers. Big pharmaceutical companies, such as Bristol-Myers Squibb and Novartis, have configured stand-alone solutions that, by using idle time of thousands of desktop computers (cycle-scavenging), acquire teraflops of cheap computing power. Instead of stand-alone solutions, high performance computing is more easily accomplished by use of global Grid computation, which is accessible via VLs.

Drug discovery knowledge space

Information and knowledge is diverse throughout the drug discovery cycle. From target-finding to clinical trials, the character of data repositories moves from being centralized to distributed and the data itself tends to get more complex, connected and heterogeneous [26]. All the data, information and knowledge generated in a drug discovery organization over time result in a knowledge space, which is the most important R&D capital of that organization. However, current drug discovery knowledge spaces are not formalized nor organized in a generic fashion. Therefore, the content of these knowledge spaces is difficult to access and combine, thus limiting data sharing, reuse and collaboration throughout the entire drug discovery cycle. This results in suboptimal exploration of the knowledge space contents by, for instance, not recognizing possible attrition in early stages of drug discovery [30,58]. An attempt to organize drug discovery knowledge space is PharmGKB [59], which is web-based and uses genotype and phenotype categories to integrate information on diseases, drug-centered pathways, drugs and genes. Registered users can submit data in a secure fashion, which can be used in federation with data already present in PharmGKB. Although extremely helpful, the information model cannot be extended by users. To achieve a general solution for data sharing, data reuse and collaboration, VLs offer generic methods to model knowledge. Drug discovery knowledge spaces can be structured in a VL environment by employing common ontologies in combination with drug-discovery-specific ontologies [7,60].

The promise

There are no easy answers to the current problems of drug discovery. With the large-scale data generation in the life sciences by

omics biotechnologies, matters have become increasingly complicated. More details do not automatically mean more answers. At the same time, developments in ICT hold the possibility of supporting the solutions needed to get the drug discovery process afloat. A VL based on a Grid-layer (for data communication and computation), a generic VL layer with reusable methods and techniques (for information management, knowledge extraction, data analysis and data reuse) and a specific VL layer with methods and expertise to interface between the drug discovery domain and the generic VL and Grid-layer, holds the promise of a functional drug discovery PSE. This drug discovery PSE can support the virtual organizations, based on changing public–private partnerships that are a key characteristic of successful drug discovery enterprises. Moreover, it makes an integrated approach along the whole chain of the drug discovery process feasible, especially when one commits to creating a formalized knowledge space (supported by semantic web and/or Grid technology). The emerging systems or integrative biology might be instrumental for this [23,24]. At this moment, globally there are several ongoing VL and VL-like initiatives, all of which have huge financial investments and the involvement of multidisciplinary partners. However, the fair conclusion is that application of PSEs is still mostly a promise. The first initial small functional applications do not yet reflect the hundreds of millions of euros already invested. The high VL ambitions described above appear hard to substantiate. Nonetheless, it is evident that *ad hoc* solutions based on increasing numbers of isolated components become unworkable. Therefore, development of VL and VL-like systems is essential for answering the demand from life sciences in general and drug discovery in particular. At the same time, organizations dealing with drug discovery should realize that the only way to adopt the principle of collaboration via a VL–PSE is in practice [11]. Drug discovery organizations should start participating in VL initiatives so that they will learn how VLs operate. Only then will they be able to mould the drug discovery PSE to their own specific demands, thereby creating a new future for drug discovery.

Acknowledgements

This work was carried out in the context of, and with the input of, many collaborators of the Virtual Laboratory for e-Science project (www.vl-e.nl). This project is supported by a BSIK (Besluit Subsidies Investeringen Kennisinfrastructuur) grant from the Dutch Ministry of Education, Culture and Science and is part of the ICT innovation program of the Dutch Ministry of Economic Affairs. We would like to acknowledge the critical reading and linguistic suggestions of Jenny Batson.

References

1 Weinstein, J.N. (2002) 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. *Curr. Opin. Pharmacol.* 2, 361–365
2 Neumann, E. (2005) A life science semantic web: are we there yet? *Sci. STKE* 2005, pe22
3 De Roure, D. and Hendler, J.A. (2004) e-Science: the grid and the semantic web. *IEEE* 1094, 65–71
4 Hey, T. and Trefethen, A. (2003) e-Science and its implications. *Philos. Transact. A. Math. Phys. Eng. Sci.* 361, 1809–1825
5 Nilsson, T. (2003) Virtual laboratories in the life sciences. A new blueprint for reorganizing research at the European level. *EMBO Rep.* 4, 914–916
6 Afsarmanesh, H. *et al.* (2002) VLAM-G: a grid-based virtual laboratory. *Scientific Programming* 10, 173–181
7 Roos, M. *et al.* (2004) Future application of ontologies in e-Bioscience. In *W3C Workshop on Semantic Web for Life Sciences*
8 Ball, P. (1999) The speed of computers. *Nature* 402, C61
9 Gerhold, D.L. *et al.* (2002) Better therapeutics through microarrays. *Nat. Genet.* 32(*Suppl*), 547–551
10 Howbrook, D.N. *et al.* (2003) Developments in microarray technologies. *Drug Discov. Today* 8, 642–651
11 Roses, A.D. *et al.* (2005) Disease-specific target selection: a critical first step down the right road. *Drug Discov. Today* 10, 177–189
12 Kaul, A. (2005) The impact of sophisticated data analysis on the drug discovery process. *Business Briefing. Future Drug Discovery* 2005, 48–53
13 Gershell, L.J. and Atkins, J.H. (2003) A brief history of novel drug discovery technologies. *Nat. Rev. Drug Discov.* 2, 321–327
14 Kramer, R. and Cohen, D. (2004) Functional genomics to new drug targets. *Nat. Rev. Drug Discov.* 3, 965–972
15 Booth, B. and Zemmel, R. (2004) Prospects for productivity. *Nat. Rev. Drug Discov.* 3, 451–456
16 Chapman, T. (2004) Drug discovery: the leading edge. *Nature* 430, 109–115
17 Burbaum, J. and Tobal, G.M. (2002) Proteomics in drug discovery. *Curr. Opin. Chem. Biol.* 6, 427–433
18 Buchanan, S.G. *et al.* (2002) The promise of structural genomics in the discovery of new antimicrobial agents. *Curr. Pharm. Des.* 8, 1173–1188
19 Werner, E. (2003) In silico multicellular systems biology and minimal genomes. *Drug Discov. Today* 8, 1121–1127
20 Apic, G. *et al.* (2005) Illuminating drug discovery with biological pathways. *FEBS Lett.* 579, 1872–1877
21 Clark, D.E. and Newton, C.G. (2004) Outsourcing lead optimisation–the quiet revolution. *Drug Discov. Today* 9, 492–500
22 Verkman, A.S. (2004) Drug discovery in academia. *Am. J. Physiol. Cell Physiol.* 286, C465–C474
23 Butcher, E.C. *et al.* (2004) Systems biology in drug discovery. *Nat. Biotechnol.* 22, 1253–1259
24 Hood, L. and Perlmutter, R.M. (2004) The impact of systems approaches on biological problems in drug discovery. *Nat. Biotechnol.* 22, 1215–1217
25 Schadt, E.E. *et al.* (2003) A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem. Soc. Trans.* 31, 437–443
26 Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4, 45–58
27 Chien, A. *et al.* (2002) Grid technologies empowering drug discovery. *Drug Discov. Today* 7(*20 Suppl*), S176–S180
28 Falk Hoffman, H. (2004) *Statement from CERN and the scientific Community*, European Organization for Nuclear Research
29 Peitsch, M.C. *et al.* (2004) Informatics and knowledge management at the Novartis Institutes for BioMedical Research. *Scip.online* 46, 1–4
30 Hug, H. *et al.* (2004) Ontology-based knowledge management of troglitazone-induced hepatotoxicity. *Drug Discov. Today* 9, 948–954
31 Bard, J.B. and Rhee, S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* 5, 213–222
32 Canessa, E. *et al.* (2002) Virtual laboratory strategies for data sharing, communications and development. *Data Science Journal* 1, 248–256
33 Walker, D. *et al.* (2000) The software architecture of a distributed problem-solving environment. *Concurrency: Practice and Experience* 12 (15)
34 Cannataro, M. *et al.* (2004) Proteus, a grid based problem solving environment for bioinformatics: architecture and experiments. *IEEE Computational Intelligence Bulletin* 3, 7–17
35 Schuchardt, K. *et al.* (2002) Ecce-a problem-solving environment's evolution toward Grid services and a Web architecture. *Concurrency Computation. Practice and Experience* 14, 1221–1239
36 Barnatt, C. (1990) Office space, cyberspace & virtual organization. *Journal of General Management* 20, 78–91
37 Foster, I. *et al.* (2001) The anatomy of the grid. *Int. J. High-Performance Comput. Applic.* 15, 200–222
38 Foster, I. *et al.* (2002) The physiology of the grid: an open grid services architecture for distributed systems integration. in *Open Grid Service Infrastructure Working Group*. Global Grid Forum
39 Stevens, R.D. *et al.* (2003) myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 19(*Suppl 1*), i302–i304
40 Chervenak, A. *et al.* (2000) Towards an architecture for the distributed management and analysis of large scientific datasets. *J. Netw. Comput. Appl.* 23,

187–200
41 Cannataro, M. and Talia, D. (2003) The knowledge grid. *Commun. ACM* 46, 89–93
42 Hewett, M. *et al.* (2002) PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res.* 30, 163–165
43 Stevens, R.D. *et al.* (2004) myGrid and the drug discovery process. *Drug Discov. Today. BIOSILICO* 2, 140–148
44 Clark, T. *et al.* (2004) Globally distributed object identification for biological knowledgebases. *Brief. Bioinform.* 5, 59–70
45 Hodgman, C. (2001) An information-flow model of the pharmaceutical industry. *Drug Discov. Today* 6, 1256–1258
46 Carel, R. and Pollard, J. (2003) *Knowledge Management in Drug Discovery R&D* 3re Millennium Inc.
47 Potter, J.D. (2001) At the interfaces of epidemiology, genetics and genomics. *Nat. Rev. Genet.* 2, 142–147
48 Buetow, K.H. (2005) Cyberinfrastructure: empowering a "third way" in biomedical research. *Science* 308, 821–824
49 Sanchez, W. *et al.* (2004) *caGRID White Paper* National Cancer Institute Center for Bioinformatics NCICB
50 Covitz, P.A. *et al.* (2003) caCORE: a common infrastructure for cancer informatics. *Bioinformatics* 19, 2404–2412
51 Martone, M.E. *et al.* (2004) E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat. Neurosci.* 7, 467–472
52 Watson, C. (2004) Carol Goble discusses the impact of semantic technologies on the life sciences. *Drug Discov. Today BIOSILICO* 2, 4–6
53 Knowles, J. and Gromo, G. (2003) A guide to drug discovery: Target selection in drug discovery. *Nat. Rev. Drug Discov.* 2, 63–69
54 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262–275
55 Lengauer, T. *et al.* (2004) Novel technologies for virtual screening. *Drug Discov. Today* 9, 27–34
56 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055
57 Claussen, H. *et al.* (2004) The FlexX database docking environment - rational extraction of receptor based pharmacophores. *Current Drug Discovery Technologies* 1, 49–60
58 Hug, H. *et al.* (2004) ADRIS - the adverse drug reactions information scheme. *Clin. Neuropharmacol.* 27, 44–48
59 Altman, R.B. *et al.* (2003) Indexing pharmacogenetic knowledge on the world wide web. *Pharmacogenetics* 13, 3–5
60 Hendler, J.A. *et al.* (2002) Integrating applications on the semantic web (English version). *Journal of the Institute of Electrical Engineers of Japan* 122, 676–680