INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

OpenMolGRID

PROJECT PLAN

Contract Reference:	**IST-2001-37238**
Document identifier:	**OpenMolGRID-WorkPlan**
Date:	**16/02/2004**
Work package:	**WP7: Project Management**
Partner	**FZJ, UT, CGX, OMC, UU, NEGRI**
Lead Partner	**FZJ**
Document status:	**APPROVED**
Classification:	**PUBLIC**
Deliverable identifier:	**D7.1a**

Abstract:
The Project's Work Plan describes the project's objectives and the work to be carried out.

Project information

Project acronym:	OpenMolGRID
Project full title:	Open Computing GRID for Molecular Science and Engineering
Proposal/Contract no.:	IST-2001-37238
European Commission:	
Project Officer:	Annalisa BOGLIOLO
Address:	European Commission - DG Information Society F2 - Grids for Complex Problem Solving B-1049 Brussels Belgium
Office:	BU31 4/79
Phone:	+32 2 295 8131
Fax:	+32 2 299 1749
E-mail	annalisa.bogliolo@cec.eu.int
Project Coordinator:	Mathilde ROMBERG
Address:	Forschungszentrum Jülich GmbH ZAM D-52425 Jülich Germany
Phone:	+49 2461 61 3703
Fax:	+49 2461 61 6656
E-mail	m.romberg@fz-juelich.de

Contents

1. Project Summary

Objectives

The specific objectives of this project are as follows: (1) To use EUROGRID for integrating heterogeneous and distributed databases for computational molecular engineering, (2) To use EUROGRID for integrating relevant existing tools for carrying out molecular modelling, (3) To provide a solid foundation for the design of next generation molecular engineering tools (prediction of molecular structures with target properties), (4) To provide secure global access to highly sensitive R&D information resources through EUROGRID infrastructure, (5) To promote the use of the OpenMolGRID environment for scientific and industrial end-users, and (6) To carry out representative tests for global life science applications.

Description of the Work

The OpenMolGRID project will address large-scale molecular design problems. The molecular design in a large is essentially based on data warehousing and data mining. Data warehousing techniques are needed to collect relevant data from distributed and heterogeneous databases. Data mining techniques (MLR, PCA, PLS, ANN, etc) are used to build predictive QSPR/QSAR models. The Grid approach is vitally essential, since the amount of data is huge and model building is computationally very demanding. The first step to address these issues is to use EUROGRID infrastructure for adaptation of existing software modules, and to make solid foundation for next step molecular engineering tools. This involves the design of seamless and unified user interface and providing adapters to make software grid-aware. The second step is to use OpenMolGRID system to develop prototype application for generation molecular structures with given chemical properties or biological activities. Thirdly, the OpenMolGRID will involve intensive testing of developed system by academic and industrial users on real application (multi-drug resistance, GPCR activity, and toxicity).

Milestones and Expected Results

The expected milestones/results are as follows: (1) grid-aware tools for accessing heterogeneous and distributed databases, (2) grid-aware tools for developing predictive QSPR/QSAR models, (3) software prototype(s) for automatic computational molecular engineering (generating structures with predefined chemical properties or biological activities).

2. Project Objectives

Molecular engineering is the task of designing molecular compounds and materials with predefined target properties. The challenge in the industrial application of molecular engineering is to design compounds that up to present have not been discovered for the intended purpose and can be patented. The design of molecular compounds relies on the knowledge that the properties of molecular compounds are determined by the properties of the molecular fragments and their interaction. Molecular modelling makes use of this fact by building candidates for chemical compounds with predetermined target properties from appropriate fragments according to established rules. For all generated candidates the target properties are estimated by the quantitative structure-property relationships (QSPR), or quantum-chemical modelling. Finally, candidates that match the predefined target property are selected for laboratory tests (see also Figure 1 in Work Plan: General Description).

The QSPR relies on the observation that molecular compounds with similar structure have similar properties. For each specific application a set of molecules is needed for which the target property is known. This requires a search of globally distributed information resources for appropriate data. For the purpose of exploring molecular similarity, descriptors are calculated from the molecular structure. Thousands of molecular descriptors have been proposed and are used to characterise molecular structures with respect to different properties. Their calculation puts high demands on computer resources and requires high-performance computing. For the available set of compounds with the appropriate target property a model for the QSPR is developed. This involves finding the most suitable theoretical method and set of descriptors. Finally, the developed model is used to predict the properties for the new molecular compounds.

The main objective of the proposed OpenMolGRID project is to provide a unified and extensible information-rich environment for solving molecular design/engineering tasks relevant to chemistry, pharmacy and life sciences. This will be achieved by extending the currently used local approach to the global dimension by building the OpenMolGRID environment on top of the Grid infrastructure provided by the EUROGRID project. The planned system will provide seamless integration of existing, widely accepted, relevant computing tools and data sources. The proposed system will target both academic and commercial end-users (especially chemical and pharmaceutical industry).

The OpenMolGRID system will comprise a set of application-oriented tools that are built on core Grid services and functions provided by the EUROGRID infrastructure. The specific objectives of the project are as follows:

- To develop tools that permit end-users to securely and seamlessly access, integrate, and use globally distributed information resources and systems relevant to molecular engineering.
- To develop tools that permit end-users to securely and seamlessly access, integrate, use, and schedule globally distributed computational methods and tools used for molecular engineering.
- To provide a realistic test bed and reference application for similar Grid projects in life science and beyond.
- To promote the use and evolution of both the EUROGRID and OpenMolGRID environment for scientific and industrial end-users.
- To provide foundations and design principles for developing and constructing next-generation molecular engineering systems.

Secure and seamless access/integration of distributed resources

The accurate and reliable prediction of chemical structures with pre-defined physical, chemical, optical, electrical or thermodynamic properties and/or biological activity draws on the known properties of chemical compounds. The accuracy and reliability of the predicted molecular structures critically depends on the amount of information accessible for analysis.

Huge amounts of information have already been accumulated worldwide. Modern molecular engineering requires that relevant portions of the available data is located, filtered, combined, and transformed in order to be used by subsequent analysis and modelling processes. The project will develop tools that permit end-users to securely and seamlessly locate, integrate, and use this information.

In addition to information resources, computational molecular modelling and molecular engineering involve a large variety of methods and software based on quantum and statistical mechanics, chemical and bioinformatics and other theoretical approaches. Using complementary techniques, the relevant software has been developed and made available by laboratories throughout the world. To facilitate the design of new compounds, molecular engineers need to make efficient use of those methods and computational resources. The project will develop tools that permit molecular engineers to securely and seamlessly access, configure, schedule, and execute the relevant programs and integrate the results.

To develop the tools required to use and integrate the relevant information and computational resources, the project will develop the necessary concepts and software on the basis of the existing EUROGRID infrastructure.

Test bed and reference application for life science

A test bed for the OpenMolGRID will be developed and evaluated in co-operation with a high-profile software company specialising in life science. End-user testing and evaluation will be carried out in conjunction with international chemical and pharmaceutical companies.

Promotion of the use of the OpenMolGRID and UNICORE

Grid computing is still a very young discipline. The ultimate purpose of it is to empower users and organisations to work effectively in an information-rich environment. The UNICORE provides an infrastructure for applications like OpenMolGRID. Providing a Grid-based environment for theoretical and computational molecular science and molecular engineering, the OpenMolGRID will help to demonstrate the power of the general Grid concept and EUROGRID's UNICORE in particular.

Foundation for designing next-generation molecular engineering tools

The ultimate goal of molecular engineering is the accurate and reliable prediction of chemical structures with pre-defined properties (e.g., physical, chemical, optical, electrical, thermodynamic, or biological activity). Such predictions are of extreme interest for chemical and pharmaceutical industries because they will allow the development and patenting of new generic compounds and materials. The OpenMolGRID will facilitate the foundations for developing next-generation molecular engineering tools.

Success criteria:

During the project the success of the OpenMolGRID project will be measured against the following criteria:

- Integration of the relevant information and computational resources into OpenMolGRID
- Integration of the tools to use the information and computational resources made available in OpenMolGRID
- Design, implementation and prototype implementation of an end-user control panel for OpenMolGRID
- Evaluation of the OpenMolGRID by industrial end users

Ultimately, the success will rely on the incorporation of the OpenMolGRID into the European Grid infrastructure and its use in E-Science by industrial end-users.

3. Participant List

List of Participants

Partic. Role[1]	**Partic. no.**	**Participant name**	**Participant short name**	**Country**	**Date enter project**[2]	**Date exit project**[2]
P[3]	1	University of Tartu	UT	Estonia	Start of project	End of project
P	2	University of Ulster	UU	Northern Ireland	Start of project	End of project
P	3	Mario Negri Institute	NEGRI	Italy	Start of project	End of project
C[4]	4	Forschungszentrum Jülich GmbH	FZJ	Germany	Start of project	End of project
P	5	ComGenex, Inc.	CGX	Hungary	Start of project	End of project
A	6	OpenMolCONSULTING	OMC	Germany	Start of project	End of project
A	7	Politecnico di Milano	POLIMI	Italy	Start of project	End of project

[1] C = Coordinator ; P - Principal contractor; A - Assistant Contractor

[2] Normally insert “Start of project” and “End of project”. These columns are need for possible later contract revisions caused by joining/leaving participants

[3] Coordinator from start of project to 30/11/2003

[4] Coordinator from 01/12/2003 to end of project

4. Contribution to Programme/Key Action Objectives

The OpenMolGRID project addresses Cross Programme Action V.1.9 “Grid Technologies and their applications”. In particular, the project will contribute to following CPA9 objectives:

- The project will conduct trials aiming at the introduction of the "Grid at large" in industrial, commercial and societal applications. This trial consist of two major steps: (i) integration of existing technologies and set-up of a Grid environment responding to true industrial, commercial and/or societal requirements; (ii) preparation, execution and evaluation of a number of applications driven by real users and carried out in collaboration with the respective technology and service providers.
- It will allow seamless access (data warehousing) to dispersed information, and aids knowledge discovery and extraction (data mining) from distributed knowledge resources.
- It will develop reusable components that allow interoperation both at the system and at the application level.
- It will enable future projects to benefit from the OpenMolGRID project.
- The project will use and promote standards of Grid computing for molecular design applications.
- It will address security issues by using existing standards provided by EUROGRID.
- It supports the international collaboration of researchers.
- It supports the European E-Science initiative and makes global resources more accessible to European scientists.

It makes non-European markets aware of European R&D results and promotes the resulting products.

5. Innovation

The proposed OpenMolGRID provides a unified and extensible information- and tool-rich environment for solving molecular design/engineering tasks by extending the currently used local approach to the global dimension using the UNICORE infrastructure. By embracing the Grid dimension, OpenMolGRID provides a new quality of service with respect to the information and methods required in the molecular engineering process. As a consequence, the scientific results obtained will rise to new levels of reliability and accuracy. Furthermore, the experience gained from using and evaluating OpenMolGRID will generate new insights for designing next-generation molecular engineering tools.

The OpenMolGRID project addresses several innovative areas:

- Distributed data warehousing, data mining and computing. In particular, we will demonstrate (1) the concept of a distributed data warehouse, which will include components for virtualisation of complex data transformation operations using Grid technology, and (2) how different ITs (databases, data warehouses, data mining, and complex molecular-engineering software) relevant to complex problem solving in molecular engineering interoperate using Grid-technology to form a powerful "virtual" system. Both aspects of the project have significant potential to make contributions to Grid research and development.
- Secure global access to highly sensitive information resources
- Providing new insight into designing next-generation molecular engineering tools

5.1. Distributed data warehousing, data mining and computing

The prediction of molecular properties and the molecular design/engineering of new chemical compounds requires the location, extraction, transformation, and integration of globally distributed data sets describing environmental, toxicological, pharmacological, biological, and chemical information. State-of-the-art data warehousing methodology is normally not confronted with the degree of geographic and conceptual distribution that is inherently prevailing in this scenario. An expected innovation resulting from this research will be new mechanisms for combining and integrating such information sources on the basis of existing and emerging Grid technology. Using the UNICORE infrastructure, the distributed data warehousing and data mining concept developed in the OpenMolGRID project will permit a seamless integration of globally distributed, heterogeneous, and evolving information resources of molecular structures and properties. The OpenMolGRID will provide a unified, extensible and information-rich environment for solving molecular design/engineering tasks. As opposed to the local approach, it will permit the prediction of molecular properties and the molecular design of new chemical compounds on the most recent, globally distributed information and will largely increase the scope, reliability and accuracy of molecular engineering.

Computational molecular modelling and molecular engineering involve a large variety of methods and software based on quantum and statistical mechanics, chemical and bioinformatics and other theoretical approaches. The relevant software has been developed in globally distributed laboratories, using complementary methodology. State-of-the-art data analysis and data mining techniques require the data to be analysed be present in denormalised form. Furthermore, all the methods used for analysis must usually reside locally. However, molecular engineering tools and systems are intrinsically decentralised and vastly heterogeneous in terms of their underlying input/output format and computational requirements. An expected novelty coming from the proposed research will come from exploiting underlying Grid techniques to integrate distributed methods into a "virtual" mining system. The unified and extensible tool-rich mining and computing environment to be developed in the OpenMolGRID project will add the relevant, complementary, and globally distributed data mining and quantum chemical software to the UNICORE infrastructure. As opposed to the current approach, it will permit to predict molecular properties and to design new molecular compounds by using methods available in globally distributed laboratories. Again, this will provide a new quality of collaborative research and development and results.

5.2. Secure global access to highly sensitive information resources

Chemical and pharmaceutical industry is beginning to realise that a global platform is highly desirable for the technological application of theoretical and computational molecular sciences and molecular engineering. Security for highly sensitive proprietary information sources, tools, and applications is a major concern for all developers of technological applications. The OpenMolGRID will guarantee the high levels of security required for industrial research and development by using existing security mechanisms provided by the UNICORE infrastructure.

5.3. Providing new insight into designing next-generation molecular engineering tools

The design, development and exploration of the OpenMolGRID will bring together computer scientists, professional software developers, and researchers from chemical and pharmaceutical industry, and application-oriented scientists from academia. Their expertise covers such diverse areas as Grid technology, data warehousing and mining techniques, software engineering, life sciences, drug design, environmental protection, quantum chemistry, and molecular dynamics. The interdisciplinary co-operation of these experts required to developing the OpenMolGRID will cross-fertilise all areas involved, and foster new insights and knowledge in fielding Grid applications.

6. Community Added Value and Contribution to EC policies

The ability to adapt for the rapid global transformation into innovative society will be critical for the European community to maintain and develop the employment, growth, industrial competitiveness and the living standards of its citizens.

6.1. The European dimension of the problem

The important issue in the adaptation with the innovative society is the fast and flexible integration of the potency of new information technologies into scientific research and industrial engineering. This fast integration should enable Europe to stay forefront in many strategic areas of research, development and technological innovation. The present proposed project will be aimed to develop GRID based highly effective systems for the computational molecular engineering, with the potential use in all fields of chemical, pharmaceutical and materials development. The aimed products and results will be made available for European research and industrial communities, increasing thus their competitiveness in the global economy, particularly in the high-tech fields of bio- and nanotechnology. Importantly, the participation of Partner country (Hungary, Estonia) companies and laboratories helps in the integration of their leading scientific research institutions into European community.

6.2. The European added value

The proposed research is characterised by very high level of complexity in terms of human resources, computer requirements and the need for data and knowledge. The envisioned radically new information systems and new science paradigms are based on the prevalence of a global information ecosystem, enabling radically novel types of information systems. It will be clearly impossible to solve the necessary scientific and technical problems through a single national effort. The complexity, multi- and interdisciplinarity of the approach require the participation of versatile researchers in different fields, from life sciences, chemistry and physics, and information technology. This will be achieved by mobilising the expertise and power of leading European laboratories in related areas into joint effort. Although a sufficient expertise in individual fields is available by participating groups, a synergetic association of their expertise and resources will be mandatory to achieve the aimed goals. The support from the IST Programme would enable the unique opportunity to take the international lead in the proposed important direction of global computing.

6.3. Contribution to EU policy

The present proposal is outlined in compliance with the EU Action Plan on learning in the information society and with the priorities of the Fifth Framework Work Programme in Information Society Technologies. The ideas expressed in the proposal are directly related to the Green Paper on the Convergence of the Telecommunications, Media and Information Technology Sectors, and the Implications for Regulation. Towards an Information Society Approach (COM(97)623, December 1997) and the Green Paper on Living and Working in the Information Society: People First (COM(96)389, July 1996). The solutions proposed in the proposal should help answering the questions about the convergent information systems and the regulations of new electronic media in 21st century.

7. Contribution to Community Social Objectives

The recent EU study "Society, the endless frontier. A European vision of research and innovation policies for the 21st century" (CG-NA-17-655-EN-C) determines the socio-economic issues such as job creation, health, and environmental safety as the main criteria for successful and competitive research. The present proposal has concentrated on each of these issues.

7.1. The contribution of the project to improving the quality of life and health and safety (including working conditions)

One of the main targets of the present proposal is the development of the knowledge discovery system for life sciences. This system, when developed, has direct relevance to several important medical issues connected to health and quality of life. First, it assists to create a more adequate and competent medical service, by extracting new knowledge hidden on vast amount of biomedical data and other information. Secondly, the new modelling approaches based on OpenMolGRID system open completely new prospects for the directed molecular design, especially in the development of new, more efficient and less toxic pharmaceuticals and other biomedical agents. Thirdly, the OpenMolGRID system provides seamless and unified access to very different molecular design tools, which allows scientists to work more effectively. This significantly enhances the quality of research and self-satisfaction from their work.

7.2. The contribution of the project to improving employment prospects and the use and development of skills in Europe

The OpenMolGRID provides a major step forward for science and society by making knowledge more accessible to a wider range of people with diverse social and educational background. The benefits from this new system include faster and more economic research and development, wherefore less qualified personnel can efficiently use state of the art scientific and engineering techniques. Presently, extremely highly skilled workforce is needed in many high tech areas including the biotechnology and pharmaceutical industry. The proposed OpenMolGRID system allows for unified approach in many research and engineering fields, thus permitting less trained personnel to be employed in high tech industry and creating many more workplaces.

7.3. The contribution of the project to preserving and/or enhancing the environment and the minimum use/conservation of natural resources

The OpenMolGRID system to be developed within the proposed research will, in principle, largely assist in the design of new sustainable industrial technologies that meet the requirements of environmental protection and use minimally the natural resources. In addition, the possibility to predict adequately the toxicity, flammability and other such properties of new pharmaceuticals and industrial materials allows diminishing the work of researchers and engineers in the hazardous conditions. In addition, economically efficient environmental technologies tend to affect employment favourably. The adequate modelling of biochemical and biomedical processes by using the OpenMolGRID allows to reduce significantly the need for animal testing in biomedical research that is a major concern of environmental policies in Europe.

8. Economic Development and S&T Prospects

Grid computing is a truly global undertaking. Projects in this field can only prosper and gain credibility if they are promoted and exposed internationally. All partners understand the importance of an active information policy with respect to it and are fully committed to engage in the dissemination and exploitation of the OpenMolGRID objectives, its approach to achieve those objectives, results and technology within specific target audiences in Europe and worldwide. The main objectives of the information plan are to create awareness of OpenMolGRID, to promote taking-up of grid computing among end-users in the field of scientific computation both in academia and industry, and to exploit commercially OpenMolGRID results and expertise. The dissemination of results will be performed through different channels:

The professional community in large will be informed about the results of the OpenMolGRID project through publications in the relevant scientific journals and through presentations at research conferences and workshops. The Grid community in particular will be kept abreast of OpenMolGRID by the participating in the Global Grid Forum. The take-up of Grid computing in general will be fostered by an extensive Web presence of OpenMolGRID.

The OpenMolGRID project has significant commercial impact by opening completely new prospects for molecular design. The ability to process large data sets with state of the art molecular design tools allows to achieve a breakthrough in the development of new, more efficient and less toxic pharmaceuticals and other biomedical agents. Success in these areas will be of major strategic impact by enhancing the European competitiveness in worldwide markets. Special effort will be taken to foster the strategic impact of OpenMolGRID by disseminating the results of the project to a wide audience of end-users in chemical, pharmaceutical, and life-science industry by presentations and demonstrations at important conferences and trade shows relevant to the field. The presentation of the Grid-aware implementation of successful commercial software and the demonstration of its enlarged scope and reliability will play a key role in exploiting OpenMolGRID technology and results commercially. The direct presentation of OpenMolGRID to the international customer basis of the commercial partner will serve as the most effective way of accelerating the strategic impact of OpenMolGRID world-wide and of exploiting its results.

9. Workplan

9.1. General Description

The project work plan of the OpenMolGRID project is organised into seven workpackages. They are structured around OpenMolGRID's generic operating scenario, system processes and components as described in Section 2: Project Objectives and depicted in Figure 1. To better understand the reasoning behind the work plan structure and workpackage organisation, we describe one realistic drug design scenario. The task for this scenario is to find a set of molecular structures with given target properties, where the activity (IC50) of HIV-1 protease inhibitors is higher than X, solubility is higher than Y, and carcinogenicity is lower than Z. The process flow is depicted in Figure 1.

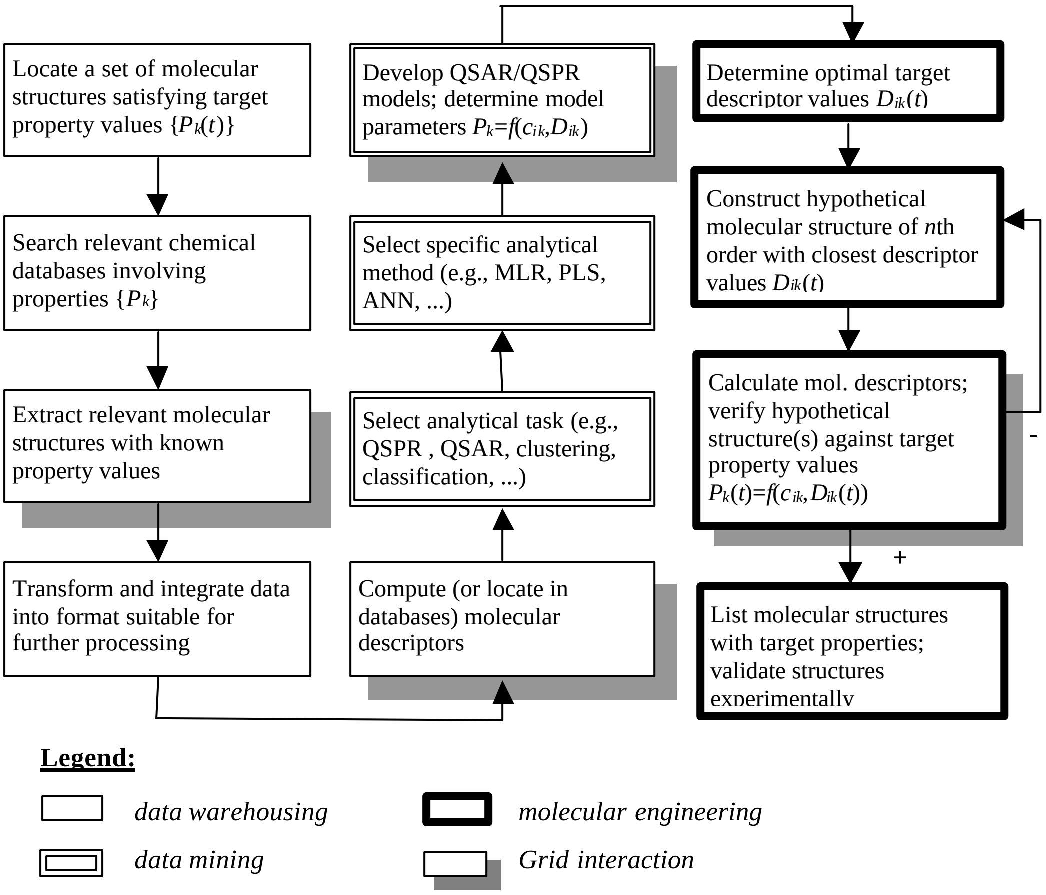
Locate a set of molecular structures satisfying target property values $\{P_k(t)\}$
Search relevant chemical databases involving properties $\{P_k\}$
Extract relevant molecular structures with known property values
Transform and integrate data into format suitable for further processing
Develop QSAR/QSPR models; determine model parameters $P_k=f(c_{ik},D_{ik})$
Select specific analytical method (e.g., MLR, PLS, ANN, ...)
Select analytical task (e.g., QSPR , QSAR, clustering, classification, ...)
Compute (or locate in databases) molecular descriptors
Determine optimal target descriptor values $D_{ik}(t)$
Construct hypothetical molecular structure of *n*th order with closest descriptor values $D_{ik}(t)$
Calculate mol. descriptors; verify hypothetical structure(s) against target property values $P_k(t)=f(c_{ik},D_{ik}(t))$
-
+
List molecular structures with target properties; validate structures experimentally
Legend:
data warehousing
molecular engineering
data mining
Grid interaction

Figure 1: Flowchart for typical drug design problem.

The flow chart in Figure 1 shows three sub-processes that need to be carried out to determine the molecular structures that exhibit the desired properties: (1) the data warehousing process (depicted by single line boxes), (2) the data mining process (double line boxes), and (3) the molecular engineering process (thick, single line boxes). Those processes that rely heavily on accessing and using computational services (high-performance, high-throughput) are highlighted by a shadow. The Grid application of infrastructure for those elements that require a seamless, reliable, secure, and efficient operation brings up a new quality in molecular engineering. This scenario is representative for a wide range of molecular design tasks. The processing structure in Figure 1 leads to a natural breakdown of the work into the following workpackages:

- WP1: Grid Data Warehousing of Molecular Structure -- Property (Activity) Information, Custom Data Repository
- WP2: Molecular Descriptor Generation and QSPR Model Building on the Grid
- WP3: Computational Molecular Engineering of New Compounds and Materials
- WP4: Grid Integration
- WP5: Test Application of the OpenMolGRID System for Chemical and Pharmaceutical Predictions
- WP6: Information Dissemination
- WP7: Project Management

The OpenMolGRID uses the Grid infrastructure as developed by the EUROGRID project (http://www.eurogrid.org) as a basis and builds on the services it offers as it is. The software system being developed by EUROGRID and a German national project is called UNICORE (Uniform access to Computer Resources). It will become freely available under the Sun community license for research by the UNICORE Forum e.V. (http://www.unicore.org).

The Grid infrastructure uses standards that are established by the Global Grid Forum (GGF, www.globalgridforum.org). The GGF is a community-initiated forum of individual researchers working on different Grid projects. The European projects UNICORE Plus (Germany, bmb+f 01-IR-001), EUROGRID (IST-1999-20247), GRIP (IST-2001-32257) as well as DataGrid (IST-2000-25182) and GRIDSTART (IST-2001-34808) are involved in the work through members who participate in different areas, research, and work groups. The GGF discusses the state of the art technology, figures out best practices, and proposes standards that are submitted to IETF. In particular, GGF deals with following areas:

- Grid Information Services, dealing with Object Specification, Notification Framework, Metacomputing Directory Services, and Relational Database Information Services
- Security, dealing with Security Infrastructure and Certificate Policy
- Scheduling and Resource Management, dealing with Advanced Reservation, Scheduling Dictionary, and Scheduler Attributes
- Performance, dealing with Monitoring Architecture and Network Monitoring
- Architecture, dealing with JINI (Java Interoperability Network Interface), Open Grid Service Architecture, Protocol Architecture, and Accounting Models
- Data, dealing with GridFTP, Data Replication, and Persistent Archives
- Applications, Programming Models & Environments, dealing with Application & Test beds, User Services, Computing Environments, Advanced Programming Models, and Advanced Collaborative Environments.

From the current taxonomy of Grid architectures the UNICORE infrastructure is a stove-piped solution covering all the layers of the Grid architecture, while for example, the Globus tool kit covers the lower level layers. As an advantage, the UNICORE infrastructure can be adapted to new applications through its plugin mechanism without having modifications to the application software. On the other end, the target system interface can easily adapted to integrate new targets.

The advantage of using UNICORE infrastructure technology is in the availability of security mechanisms, seamless access to distributed resources and flow control. It provides basic mechanisms to access distributed data sources, which can be fully exploited in WP1 to provide seamless access to relevant databases. The UNICORE security mechanisms provide single sign-on capability, authentication and authorisation as well as secure data transfer, which are important for the users as well as database providers especially from pharmaceutical companies. The ability to distribute calculations on big data sets will shorten the turn around time. These features speed up the development process and provide reliable base structure for the whole process.

The UNICORE Grid infrastructure uses state of the art Grid techniques like X509v3, Secure Socket Layer (SSL), and Public Key Infrastructure (PKI) for security and application specific interface plugins. Java makes it easy to achieve portability; it is object-oriented and platform independent. The project uses XML for information storage and exchange.

The close collaboration with the Global Grid Forum is essential for the progress of the OpenMolGRID project. Jülich will feed back developments from the OpenMolGRID to the GGF through its already established channels to different research and work groups. The OpenMolGRID project will closely collaborate with the EUROGRID project and especially with the BIOGRID workpackage of it. An agreement between the two steering committees of the two projects is planned for mutual exchange of developments.

9.2. WP1: Grid Data Warehousing of Molecular Structure -- Property (Activity) Information

The OpenMolGRID analysis process requires the location, extraction, transformation (e.g. descriptor calculation), and *integration* of globally distributed data sets describing environmental, toxicological, chemical, biological, and pharmacological information. In the first instance we will integrate data from two underlying molecular or chemical repositories, National Toxicology Program (NTP) and Ecotox. A key to understanding the OpenMolGRID data warehousing repository is that it does not integrate information to provide a "general" repository for querying molecular information, although this will be a useful by-product. Its main purpose is to provide "pre-computed" data in order to improve the efficiency and effectiveness of subsequent automatic and semiautomatic data analysis and data mining operations. By providing "cached computations" of frequently used and computationally expensive data processing tasks, it offers value for all OpenMolGRID users now and in the future (a much wider community is expected).

Under this Project Plan, we put strong emphasis on Grid and data warehousing/ management aspects of the project. To realize this we devote efforts to the following aspects:

- *Warehouse content quality and added-value*. To ensure that the data warehouse content will be useful not only for OpenMolGRID, but also to a wider molecular engineering community, we will take extra care in ensuring that the warehouse content is of interest to chemical engineers and researchers and developers in related areas. To fulfil this, we will also implement a much more sophisticated metadata approach (by implementing the Object Management Group's Common Warehouse Metamodel - CWM) and provide supplementary fields. This work is mainly conducted between UU, NEGRI, and UT.

- *Substructure search*. The data warehousing concept in OpenMolGRID includes substructure search capabilities. This function is important for identifying the best subset of data (chemical compounds) to be used for further analysis and is fundamental in chemical and related communities. Substructure searching is essentially a two-part process and results in the need for two different queries. The first query aims to select a subset of structures that may contain the substructure. This is carried out using a fingerprinting approach. Fingerprints of the structures are matched against the fingerprint of the substructure and those structures that cannot possibly match are removed from the set of matches. The second query is performed within the matching subset to select structures that actually contain the full chemical substructure. The comparison in the second step is computationally expensive. Therefore, the first step is used to remove the structures that could not match at all in the second query. To improve the performance of substructure search the data warehouse will adopt the "fingerprint" approach outlined above. This requires the fingerprint data to be stored in the warehouse for chemical structures. Special fingerprint generation and substructure programs must be invoked remotely via OpenMolGRID. This process will illustrate the ability to use other Grid resources in a distributed data warehouse solution. Work on this aspect will mainly involve UU, CGX, UT, and, for Grid aspects, FZJ.

- *Molecular descriptors*. The data warehouse will also contain a small set of molecular descriptors, which are derived via computationally intensive descriptor calculation programs, from the molecular structure information of the stored compounds. From a data warehouse perspective, these descriptor calculations could be viewed as complex data transformations. As descriptor calculation programs are often specialised in terms of required hardware and software, it is desirable to use these programs as an external resource. This scenario will be realised by using descriptor calculation programs provided by UT and CGX. Essentially, this amounts to virtualisation of parts of the data warehouse's data transformation processes. The virtualisation functionality is realized by UNICORE's the development of a command line client (i.e. task 4.6). Work on this aspect will involve partners from mainly UU, CGX (descriptor calculations programs), and for Grid aspects, FZJ.
- *Custom data repository*. The OpenMolGRID data warehouse provides the results of frequently used computations (based on certain data fields) necessary for subsequent data mining operations as *stored* data. This has the advantage that this data does not need to be computed within the data mining environments. Especially, when these data items are needed very frequently or if their computation is very time-consuming (e.g. quantum-chemical descriptors) a data warehouse solution is ideal. However, there are numerous situations where data needed for the data mining step is not obtainable from the data warehouse, for example, (1) for data items (fields, i.e. certain molecular descriptors or properties) that are only rarely used, (2) items obtained in from sources (databases, knowledge and simulation systems) that are not part of the data used to populate the data warehouse, (3) temporary items (calculation and modelling results) generated in the ongoing mining process. , and (4) any permanent data items that are generated during the use of the OpenMolGRID system (e.g. generated structures, predicted properties, QSAR models, etc.). For these cases it is absolutely necessary to have a data repository to manage the involved data. We call this system component *custom data repository* (CDR) and develop it as part of WP1. The main partner involved in carrying out the necessary work is CGX.

Collectively, the relevant processes, components, and systems constitute *data warehousing* (see warehousing subprocess in Figure 1). Currently, the computational chemistry community use a range of systems, tools, and algorithms for preparing and combining information relevant to calculating molecular properties.

The objective of this workpackage is to use the UNICORE infrastructure to integrate resources, systems and tools to develop a Grid data warehouse. To achieve the objective of the workpackage, the work will be carried out in two phases:

Phase 1 will analyse and specify the logical and technical requirements for both integrating the various utilities and tools for data warehousing and their integration and interoperation within UNICORE resulting in an overall design for a Grid Data Warehouse. This phase will require close liaison between UU and FZJ.

Phase 2 concentrates on the development and testing of the required software components arising from the requirements specification. This phase will be characterised by a highly iterative, rapid-prototyping approach to build and maintain project momentum, help identify design inadequacies, and allow for quicker deployment (i.e., UNICORE technical integration, end-user evaluation).

The work of WP1 will be carried out in two major steps, corresponding to two tasks: requirements analysis and specification (Task 1.1), and design, implementation, and testing of software components (Task 1.2).

Task 1.1: Analyse and Specify Requirements

A thorough requirements analysis and specification will be conducted to identify and describe the required functionality for integrating the data, information, and computing components needed for a seamless interoperability and scheduling of these resources and services. This will enable a design for the data warehousing components to be formulated. This task will require intensive collaboration between UU and FZJ. It will also require input form UT and NEGRI. Key elements of the specification will include:

- the analysis and specification of a common data input/output format (file) used by data transformation methods including appropriate adapters/converters for the various warehousing methods; the adapters must be developed in accordance with UNICORE requirements for data transfer and communication.
- analysis and the specification of complementary data warehousing and integration methods that allow a flexible combination and adaptation of existing data processing methods currently used by the computational chemistry community.
- analysis and the specification of logical flow of the warehousing process and the scheduling of the involved physical resources (e.g., data repositories and processors).

Task 1.1 can be broken down into subtasks. The described specification elements are distributed between these subtasks.

Task 1.1.1 - Understanding, Requirements and High-Level Architecture

This task is mainly concerned with getting to understand the problem in context and analysing the user requirements resulting in specifications being developed. A high-level architecture design for the data warehouse, including its interfaces with UNICORE, will be developed as a result of these specifications. In addition the need for metadata will be analysed and specified as appropriate. These specifications will be presented in various deliverables.

Task 1.1.2 - Logical Model, Transformations and Physical Model

A logical data model is required to identify the structure that data will be presented with the data warehouse. This will be developed using the standard database approach that will be refined to suit the data warehousing process. Once the logical model is developed, derived fields and transformations can be defined, enabling the physical model to be developed. This task will focus on these processes. This task is largely dependant on the output of Task 1.1.1.

Task 1.1.3 - Data Source Analysis

Data sources will be analysed with respect to the logical model to determine if the necessary data is available within these sources, or if other transformation are required. This task will analyse each individual data source identified, to determine if it is suitable for integration according to the model developed during task 1.1.2.

Task 1.1.4 - Specification of Custom Data Repository (CDR)

Some data required by subsequent processes can be considered as temporary or non-stable data and as such cannot be dealt with by the data warehouse. On the other hand, many data items are generated during the data mining process when e.g. developing QSAR/QSPR models, predicting properties or enumerating chemical structures that need to be stored as permanent data for later use. Such data is in violation to its fundamental concepts of data warehousing and this introduced the need to develop a CDR to store this data. In addition, the CDR will be used as input to the data warehousing process to illustrate that potentially proprietary data sources can be integrated into the data warehouse. Task 1.1.4 aims to specify the requirements for the CDR.

A detailed list of requirements for the resulting components will be generated in the specification process. The specification will mainly be Deliverable D1.1a and is scheduled for project month 5. Other deliverables will support this deliverable and will be available at various points during the project.

Task 1.2. Design, Implementation, and Testing of Developed Components

Depending on the specific contents of D1.1a and other deliverables, the design and implementation task will comprise:

- the design, implementation, and testing of the data warehouse, its data structures and their method-specific adapters for the data warehousing process on the UNICORE infrastructure.
- the design, implementation, and testing of warehousing-specific methods not yet available used by the computational chemistry community.
- the design, implementation, and testing of the CDR.

This task can be broken down into several subtasks. The described phase is distributed between these subtasks.

Task 1.2.1 - Command Line interaction for descriptor calculation, fingerprinting and Substructure search

This task aims to enable the data warehouse to act as a UNICORE client by interacting with a command line tool. he need to act as a client comes from the fact that a number of calculations may greatly reduce the ability to respond to requests efficiently, especially in relation to descriptor calculation, fingerprinting and substructure searching. This task aims to use the command line interface to enable these calculations to be carried out in a distributed manner. Currently a command line interface does not exist within UNICORE and will be developed as part of WP 4.

Task 1.2.2 - Design of the data warehouse and Common Processing Environment, Parsing tools, and CWM components.

The analysis of data sources will result in the identification of information to be extracted from the individual sources and the subsequent processing to be carried out on this information to make it suitable for inclusion into the data warehouse. Some data processing and transformations will be common among all data sources, leading to the need to develop a Common Processing Environment or CPE for data processing. This task will result in the design of the CPE. In addition, the physical model will be described in terms of the CWM in order to provide metadata.

Other data processing and transformations will be specific to a particular data source, resulting in the need for an individual parser to be developed for each data source. Based on the CPE, the specific data processing for each source will be designed and based on the CWM. At this point care must be taken in order to reduce code duplication in subsequent steps. This task is dependant in part to Task 1.2.1.

Task 1.2.3 - Implementation and Testing of the data warehouse, CPE, Parsing Tools and CWM

The purpose of this task is to implement the data warehouse and the tools necessary for it to function correctly including the CPE and parsing tools for each data source. Once the CPE and parsing tools have been designed, these can then be implemented. This task is therefore dependant on Task 1.2.2.

Task 1.2.4 - Implementation of CDR

The purpose of this task is to implement the CDR and the interface surrounding it to make it available as a Grid resource. This task is therefore dependant on Task 1.1.4. This implementation involves the design and preparation of a relational database and the development of an interface for managing the CDR.

The integration of the developed components into the UNICORE Infrastructure will be carried out in WP4, and the end-user testing will be performed in WP5. Task 1.2 will result in Deliverable D1.2 and Deliverable D1.5, which consists of a set of software components (and their documentation), which implement various parts of the OpenMolGRID warehousing processes (see Figure 1). NEGRI and CGX will provide several important data sources required for the development and commercial exploitation of OpenMolGRID. In addition, public data sources like National Toxicology Program (NTP), and Ecotox will be accessed.

9.3. WP2: Molecular Descriptor Generation and QSPR Model Building on the Grid.

The main target of this workpackage is the adaptation of multiple existing application software for the QSPR/QSAR analysis to be executed in the Grid environment. A variety of theoretical methods and computer software is available of theoretical prediction of chemical properties and biological activity of compounds based on designated experimental data sets on target properties and the corresponding theoretically calculated molecular descriptors. On the top of the Grid environment, a common interface for different domain specific applications will be provided to make job submission easier and more transparent to the end users. Many of the available methods are very demanding in terms of computer hardware and execution time. However, the Grid environment can be used for selection and aggregation of distributed resources across multiple organisations for solving large scale computational and data intensive problems.

Task 2.1: Requirements specification.

The requirements for integrating the specific QSPR/QSAR software into the UNICORE infrastructure will be analysed. The most optimal design of predictive models requires a combined application of multiple software packages (e.g. different programs for molecular descriptor calculation, data analysis, and model building). The effective use of different applications is not a trivial task, because different programs are generally incompatible with each other. Thus, the development of unified and application neutral formats for describing input and output data of QSPR/QSAR software tools will be developed. This workpackage requires close cooperation with WP1 and WP4 to achieve optimal interaction between different modules and the UNICORE infrastructure. All other tasks in WP2 will use the outcome of this task

Task 2.2: Adaptation of molecular descriptor calculation software.

A variety of complex software is available for the calculation of molecular descriptors. For a single compound, thousands of molecular descriptors can be computed directly from the information encoded in its structural formula and the relevant background information. Initially, the descriptor calculation (MDC) module from the CODESSA package will be adapted to be universally accessible within the UNICORE environment.

Task 2.2.1 deals with the adaptation of the MDC module for the UNICORE infrastructure; Task 2.2.2 is depending from the task 2.2.1 and deals with the development of UNICORE plugin and a wrapper for the molecular descriptor calculation task.

Task 2.3: Adaptation of quantum-chemical software needed for descriptor calculation.

The descriptor calculation software often depends on quantum-chemical or molecular dynamics calculations. The respective software is available both at the *ab initio* (GAUSSIAN, GAMESS, etc.) and semi-empirical (MOPAC, AMPAC, ZINDO, etc.) level of theory. A number of molecular dynamics packages (AMBER, GROMOS, Car-Parinello, etc.) are available as well. However, the installation sites are, as a rule, distributed geographically and often deploy different hardware and software platforms. An important task evolves thus from the need to uniformly access quantum chemical software on the Grid as requested by clients calculating molecular descriptors. In addition, the development of the necessary Grid interfaces will substantially benefit from the tools (e.g. interfaces for Car-Parrinello Molecular Dynamics and Gaussian 98) developed within the active EUROGRID project. For the OpenMolGRID project MOPAC software will be initially adapted for the semi-empirical quantum chemical calculations. In addition, the MOLGEO software will be adapted for the conversion of molecular structures from the 2D representation to the 3D representation. This is required because the data sources often have molecular structures in 2D representations, but the quantum chemical calculations require molecular structures in 3D representations.

Task 2.3.1 deals with the adaptation of MOLGEO software for the UNICORE infrastructure, while subsequent Task 2.3.2 develops the UNICORE plugin and wrapper for the 2D to 3D conversion task;

Task 2.3.3 deals with the adaptation of MOPAC software for the UNICORE infrastructure, while subsequent Task 2.3.4 develops the UNICORE plugin and wrapper for the MOPAC calculations.

Task 2.4: Adaptation of existing software for QSPR/QSAR model (MLR, PCA, ANN, etc.) development in Grid environment.

In this task, a number of existing QSPR/QSAR methods from the CODESSA software package will be adapted for the Grid environment. The available methods for developing QSPR/QSAR models include the multilinear regression (MLR), multivariate analysis (PCA, PLS), artificial neural networks (ANN), and other non-linear representations of the quantitative structure-property/activity relationships. Each of the above-listed methods requires specific input information with respect to chemical, physical or biological property data and the theoretically calculable molecular descriptors. Thus, this task is closely correlated with WP1 and Tasks 2.2 and 2.3 from this workpackage.

Task 2.4.1 adapts the model building module from the CODESSA package for the UNICORE infrastructure; Task 2.4.2 is depending on the previous subtask and develops the UNICORE plugin and wrapper for the QSPR/QSAR model building task.

9.4. WP3: Computational Molecular Engineering of New Compounds and Materials

The primary task of this workpackage will be the development of software for hypothetical molecular structure building with target properties and its implementation within the Grid. The QSPR/QSAR models contain information about the relationship between the structural elements of the molecule and the property values. This enables us to apply a more refined approach that takes advantage of the possibility to divide the molecule into functionally and spatially different fragments. For instance, for each fragment, a set of molecular descriptors can be assigned that in combination give the initial (0th order) approximation to the true descriptor value. By iterative adjustment of the descriptor values for hypothetical molecular structures and using structural fragment library, it will be possible to extend the search space to the potentially useful compounds that have not yet been synthesised. This process can be, however, very time consuming and require the software and/or databases distributed geographically on different sites. Such molecular engineering requires also the availability of a comprehensive database of molecular fragments, which is used for rapid estimation of the actual descriptor values. Also, the methodology for finding optimal descriptor values from the QSPR/QSAR model, and rules for the combination of different fragments to form molecules need to be developed. The development of the Grid tools and the computer software applicable for the hypothetical structure building would be an important task within this workpackage.

Task 3.1: Requirements specification.

The requirements specific to the molecular engineering applications will be analysed. The analysis covers the selection of most appropriate methods and meeting the requirements of UNICORE standards. This workpackage will use modules integrated by WP1 and WP2, and wrappers developed by WP4. All other tasks in WP3 will use the outcome of this task.

Task 3.2 Development of a fragment library for the generation of molecular structures.

A database of molecular fragments will be created within this task. The fragments will be divided into different classes depending on their role in determining the different molecular properties or biological activity. Such classes will include the generic structures and active and passive substituents towards the given property. Each fragment in the library will have a large number of molecular descriptors (up to a few thousand) to be stored in a special database. The initially targeted fragment library will include above 2,000 generic structures and more than 100 substituents that in combination allow to generate several million target molecular structures. The library will have an open structure to be extendable by any number of fragments of different types.

Task 3.2.1 specifies the data structures for the fragment library and provides initial set of data that are used for the implementation of Task 3.3, Task 3.4, and Task 3.5; In addition, the subsequent Task 3.2.2 implements the UNICORE plugin and wrapper for the fragment library.

Task 3.3: Development of a methodology for creating molecular structures with given property values.

The set of rules for the generation of a hypothetical molecule from the available structural fragments will be developed and applied within the special software. Methods for hypothetical structure generation can be evolved from the case based reasoning, structural evolution techniques, and genetic algorithms. Also, some of above mentioned methods could complement each other. Some of these algorithms are very demanding in terms of computational resources. Therefore, a necessary Grid interface will be developed to enable the most efficient implementation of various strategies for the computational molecular structure building.

Task 3.3.1 implements software for the structure generation, while the subsequent Task 3.3.2 will implement UNICORE plugin and wrapper for the structure generation.

Task 3.4 Assignment of approximated descriptor values for elements in the structural fragment library.

The methodology will be developed and implemented to carry out the rapid estimation of approximate descriptor values for hypothetical target molecular structures. This is a general optimisation task that can be solved using iterative self-consistent algorithms, artificial neural networks or genetic algorithms. The respective software developed produces a large set of hypothetical molecules.

Task 3.4.1 adapts MDC module for calculating fragment descriptors. The outcome of this task is used by the Task 3.3. 1. The Task 3.4.1 is followed by Task 3.4.2, which develop UNICORE plugin and wrapper for the calculation of fragment descriptors.

Task 3.5 Development of methods for the calculation of actual descriptor values based on fragment descriptor values.

This task takes care of the fast and reliable calculation of actual descriptor values and predicting property/activity values with the help of QSPR/QSAR models. In general, different conformations should be checked, since descriptor values significantly depend on a 3D structure of the molecule. The process will involve time-consuming quantum-chemical or molecular dynamics calculations for extensive sets of molecules, and should be built on top of the UNICORE infrastructure. This task is carried out with close cooperation with WP2.

Task 3.5.1 adapts software for calculation actual descriptor values and predicting property/activity values. This task is followed by the Task 3.5.2, which implements respective UNICORE plugin and wrapper.

9.5. WP4: Grid Integration

This workpackage will provide the user interface and the Grid services for the process of determining a set of molecule structures with specified target properties. It uses the Grid infrastructure as developed by the EUROGRID project. A major task in this workpackage is to add seamless access to different databases to the Grid infrastructure by exploiting the appropriate UNICORE interfaces. It will enhance the seamless interface for database access, new applications, and workflow by using the UNICORE plugin-mechanism. The UNICORE concept of asynchronous meta-computing together with resource brokering can be used to speed up the computational tasks. The workpackage is split into seven tasks, which correspond to the requirements for Grid integration from workpackages WP1 – WP3, to the overall process, the Grid infrastructure support, and the integration testing. Each task includes technical testing of the respective prototype.

Task 4.1: Design and implementation of database access

Access to databases as a data source is currently not provided by UNICORE but essential for the data warehousing part of the project (WP1). In this task the open interfaces of the UNICORE infrastructure will be used to develop necessary additions to the system. The plugin-mechanism in the UNICORE client allows to add new application specific functions to the graphical user interface. The requirements specification from workpackage one (WP1) is taken here as the basis for the design of a seamless, platform and application independent interface for database access. A graphical user interface will be developed and added to the existing UNICORE user interface as a plugin written in Java. The GUI requests all input from the user necessary to do the database query or upload as derived from the specifications in WP1. It includes consistency checks for the input to ensure that all necessary information is provided to the server. The interface generates the Abstract Job Object (AJO) and sends it to the selected UNICORE server. On the server side of the Grid system the uniform queries and upload requests have to be translated into requests for the target database, execute them, and the resulting output has to be made available in a format suitable for further processing. To achieve this, prototype database access tools (DBAT, the abstraction layer for database access) for selected databases will be designed and implemented. Each database will need to be accessed by a separate DBAT application, which incorporates the specific interface. The development will be iterative together with technical tests and feedback from WP1. For the initial tests UNICORE will be installed at least at two of the partner sites which provide data and compute resources.

Task 4.1.1 covers the specification of the database access tool; Task 4.1.2 the specification of the corresponding graphical user interface plugin while Task 4.1.3 and Task 4.1.4 deal with the implementation of the GUI plugin and the necessary DBATs.

Task 4.2: Design and implementation of a workflow meta-plugin and its application for descriptor calculation

This task covers the general approach for the integration of predefined workflows for calculation tasks in the molecular design and engineering process. Descriptor calculation is one of the important steps in the process. It is used in this context to prove the general concept of automated workflow support in UNICORE. The calculation of the majority of molecular descriptors requires the use of semi-empirical and ab initio quantum-chemical methods. Moreover, the descriptors have to be determined, as a rule, for a large number of molecules. Therefore, a wrapper has do be able to distribute the calculations to multiple target systems. In this task the plugin interface provided by UNICORE is enhanced to support automated workflows. A meta-plugin will be developed which supports predefined workflows and accesses application specific interfaces available in the client. These application plugins which need to interface to the meta-plugin will be developed for the necessary applications by Tasks 2.3 and 2.4. Examples are MOPAC or CODESSA Descriptor Calculation. The development will make use of the generic interface components provided by UNICORE Plus and further interfaces from BIOGRID (sub-project of EUROGRID). The meta-plugin will integrate the logic to perform asynchronous meta computing for the time-consuming calculation steps. The resource selection mechanism will make use of the resource information including software resource descriptions provided by the user accessible sites. This task is correlated to the workpackages WP1 – WP3, the requirements are jointly specified there.

Task 4.2.1 specifies the meta-plugin and its interfaces while Task 4.2.2 deals with its implementation. In Task 4.2.3 the meta-plugin is used for the descriptor calculation workflow as a first example.

Task 4.3: Development of the workflow for QSPR/QSAR model development

The data mining step comprises a set of user decisions which analytical task and method to use and of the resulting computation steps. For the integration into the Grid the tasks and methods will be made available at the partner sites as UNICORE application resources that can be used by the meta-plugin providing the automated workflow support and offered through the user interface for separate selection. The workflow to be modelled here will include support the synchronisation with the user who may need to check output and decide on data for the continuation of the workflow.

Task 4.4: Development of the workflow for the overall process

The parts developed in tasks 4.1 to 4.3 are used to model the overall OpenMolGRID process as described in the workplan: General Description (Section 9.1). This task finalizes the automated workflow interface to guide the user through the process with predefined tasks and job groups. The meta-plugin will be enhanced based on the results of the previous tasks. It will generate the steps together with its dependencies. Necessary data transfers, data conversion tasks, and synchronisation tasks are integrated.

Task 4.5: Support for the OpenMolGRID Grid infrastructure

This task defines the overall Grid architecture and infrastructure of OpenMolGRID. It establishes the OpenMolGRID testbed which comprises the partner sites. The testbed will evolve over the duration of the project related to the respective status of the components to be developed in the OpenMolGRID project.

The UNICORE security mechanism requires X.509 certificates both for users and servers. A Certification Authority (CA) which provides the necessary certificates will be established and maintained. This includes the development of an appropriate Certification Policy. Partners who will run the UNICORE infrastructure at their sites for testing the new components will be supported in installation and maintenance of the system.

Task 4.5.1 deals with the definition of the CA Policy and Task 4.5.2 with the implementation of the CA. Task 4.5.3 covers the support for the UNICORE software. All partners are installing and maintaining the OpenMolGRID testbed, this is part of Task 4.5.4.

Task 4.6: Design and Implementation of a Command Line Interface

The subject of this task is to make the data warehouse (WP 1) capable of using resources available in the Grid. This especially includes access to software resources to prepare data of important value to the data warehouse users. To accomplish this the data warehouse must be able to act as a client in the Grid. This implies the development of a command line client for UNICORE which is currently not available in UNICORE. The command line interface will be designed to generate abstract jobs, sign them with its certificate, submit them, query job status, and receive output data.

Task 4.6.1 covers the requirements analysis for the command line interface, Task 4.6.2 its specification, and Task 4.6.3 its implementation.

Task 4.7: Integration Testing

Before the OpenMolGRID system is tested with real-life applications the component developers have to combine their components and test them progressively. It is divided into two tasks: Task 4.7.1 for the integration test of database access tools and the GUI plugin and Task 4.7.2 for the integration of meta-plugin and application plugins.

9.6. WP5: Test application of the OpenMolGRID System for Chemical and Pharmaceutical Predictions

The main objective of this workpackage is to test OpenMolGRID performance in the real life chemical and pharmaceutical applications. Testing the model building capability of the GRID system in a real life application using data of human fibroblast toxicity will consist of two steps. In the first step, structure-activity model for 20,000 structures will be set up as follows: A compound library containing 20,000 diverse structures will be synthesised and assayed in vitro using human fibroblast cell lines to measure the cytotoxic effect of each of the library members. All data in the library provide information directly relevant to understanding the ADME/Tox behavior of compounds within human systems. Based on the experimental results, a model will be built by the QSPR tools of the GRID system. The accuracy of the model(s) will be validated. In the second step, the predictive capability of the model(s) constructed in the previous step will be tested against the biological activity of 10,000 previously unknown structures. In vitro human fibroblast toxicity will be experimentally determined for 10,000 newly synthesised compounds. The accuracy of the prediction by the model(s) will be validated with the comparison of the experimental and calculated results. The structure classification capability of the model, i.e. accurate selection of the most and least toxic compound sets, will also be tested.

The cost breakdown for in vitro testing is as follows:

Consumable Type	Unit price (1mg cmpd.)	Number of compounds	Amount for the period			Total
			1st half year	2nd half year	3rd half year	
Cost of Compounds for screening	12					
WP5.3		20,000	240,000	0	0	240,000
WP5.4		10,000	120,000	0	0	120,000
Subtotal			*360,000*	0	0	360,000
Cost of *in vitro* experiment	18					
WP5.3		10,000	0	137,867	42,133	180,000
WP5.4		5,000	0	68,933	21,067	90,000
Subtotal			*0*	206,800	63,200	270,000
Total			360,000	206,800	63,200	630,000

Table 1: Cost breakdown for *in vitro* tests per task and year

Cost of Compounds for screening		
Wage total	4.4	EUR/CMPD
Material cost	3.0	EUR/CMPD
Depriciation	0.5	EUR/CMPD
Overhead	4.1	EUR/CMPD
TOTAL	12.0	EUR/CMPD
Cost of in vitro experiment		
Material costs	4.5	EUR/CMPD
Labor costs	10.5	EUR/CMPD
OVERHEAD	3.0	EUR/CMPD
TOTAL	18.0	EUR/CMPD

Table 2: Cost breakdown per compound

Task 5.1: Functional testing for algorithms, modules and software frames of the Grid system as well as generate test reports.

This testing phase will investigate the general capabilities of the separate modules of the Grid system as well as solution of their integration. The following capabilities will be tested: (a) Data warehousing: accessibility, functionality of searching and effectiveness of extracting structural and property data from the system; (b) Data management: functionality of entering new data into the knowledge bases and incorporating a new database into the system; (c) Data mining: building capacity of a new QSPR model; (d) Functionality of the models for prediction or structure selection purposes; (e) Utility of the user interfaces; (f) Integration of the system.

Task 5.2: In silico testing of the Grid system for structure-activity relationship using two sets of structures having experimentally determined biological activity: Multi-Drug Resistance (MDR) and G-Protein Coupled Receptor (GPCR) activity.

Two of the most recently important activity areas in drug discovery are selected for testing purposes of the Grid system. Experimental activity values have been determined and are available for testing Grid system. The biological assays are carried out using typically small compound sets (several hundreds of structures). The purpose of the test procedure is to find the most active structures from the set. In this work, the selection capability of the Grid system will be tested through the following steps: (a) training the system for activity-structure relationship using data input of measured values available for certain structures (training set) and (b) validating selection capability of the Grid system, i.e. if the system can select the most active compounds from among other compounds (validating set) in the same library.

Description of the biological activities and their importance in the drug discovery pipeline is provided below.

Cancer chemotherapy effectiveness is limited for many patients by intrinsic or chemotherapy-induced resistance. The most common form of ***multi-drug resistance*** is the result of over-production of certain membrane proteins (MDR1/MRP1 glycoproteins), which pumps the anti-cancer drug out of the cell. MDR1/MRP1 have an important role in tissue protection at the blood-brain barrier, the blood-testis barrier, thus preventing the penetration into the central nervous system and testis.

Hundreds of compounds have been identified in a diverse discovery library, based on similarity to known ***selective GPCR ligands***. The application areas of these compounds range from neurology (depression, anxiety, epilepsy, migraine) through cardiology (hypertension, angina), metabolism (ulcers, nausea), immunology (allergies, asthma) and cancer (prostate cancer, endometrial cancer) in the case of GPCR ligand analogs, and from inhibition of tumour growth, eradication of cancer, through inhibition of proliferative diseases associated with atherosclerosis (e.g. restenosis) inhibition of angiogenesis and treatment of autoimmune diseases.

Task 5.3: Testing the model building capability of the Grid system with setting up structure-activity model for 20,000 structures.

A compound library containing 20,000 diverse structures will be synthesised and assayed *in vitro* using human fibroblast cell lines to measure the cytotoxic effect of each of the library members. All data included related to the library gives information directly relevant to understanding the ADME/Tox behaviour of compounds within human systems. Based on the experimental results, a model will be built by the QSPR tools of the Grid system. The accuracy of the model(s) will be validated.

Task 5.4: The predictive capability of the model(s) built up in the WP5.3 phase will be tested against the biological activity of 10,000 previously unknown structures.

In vitro human fibroblast toxicity will be experimentally determined for 10,000 newly synthesised compounds. Based on the experimental results, a model will be built by the QSPR tools of the Grid system. The accuracy of the prediction by the model(s) will be validated with the comparison of the experimental and calculated results. The structure classification capability of the model, i.e. accurate selection of the most and least toxic compound sets, will also be tested.

Task 5.5: Testing model testing capabilities on data available from public data sources.

This testing phase will test grid-aware data warehousing and data mining modules on different publicly available data sources (NTP, RTECS, etc). The QSAR models will be built on diverse set of compounds for several toxicological and pharmacological properties.

9.7. WP6: Information Dissemination and Exploitation of Results

A very important task for the project is the dissemination and commercial exploitation of the OpenMolGRID objectives, its approach to achieve those objectives, results and technology within specific target audiences in Europe and worldwide. The main objectives are:

- Creating awareness of OpenMolGRID;
- Promoting take-up of Grid computing among end-users in the field of scientific computation both in academia and industry;
- Commercial exploitation of results.

The dissemination and exploitation activities of the project partners are detailed in the Project Deliverable D.6.1a. As an abstract of the D6.1a, this section below summarizes the dissemination channels, the exploitation instruments, channels, methodology and schedule for such actions.

Dissemination

Grid computing is a truly global undertaking. Projects in this field can only prosper and gain credibility if they are promoted and exposed internationally. All partners understand the importance of an active information policy in this context and are fully committed to engage in the dissemination of the project results.

The dissemination will be performed through academic and public channels:

- By articles in journals;
- By presentations and demonstrations at important conferences relevant for the field of performance- and data-intensive computing;
- Through participation in the Global Grid Forum, in GRIDSTART, and in conferences or collaborations that will emerge towards the end of the project;
- With an extensive Web presence. All work will be hosted on the OpenMolGRID Web site maintained by UT with support from all partners. The prototypes developed will be available as Open or Community Source. Additional Web presence as part of the commercial exploitation is performed through CGX' and its affiliates’ web sites.

Exploitation

The OpenMolGRID technology and expertise will be exploited through all partners by the links with the other EC funded and national projects as well as the commercial partner of the project, CGX.

Exploitation Instruments

- Press Releases about to the launch of the project and achievements of evaluation versions;
- Publications and Presentations at conferences;
- The OpenMolGRID Project Summary (1 page) detailing the latest achievements of the OpenMolGRID project and Client Feedback Form (a set of questions learning more about the opinion and interest of the clients from the pharma and biotech industry and related fields on particular achievements during development of the OpenMolGRID project/system.

Exploitation channels:

- Collaborative links to other projects**:** OpenMolGRID will have collaborative links with the following EC funded and national projects. The particular projects are as follows: DEMETRA, FATEALLCHEM, EASYRING, IMAGETOX, an Italian national Project aiming a development of software to predict behaviour and toxicity of environmental pollutants.

CGX is actively serving an estimated 90%+ of the leading pharmaceutical companies worldwide. With offices in the US and Europe and representatives in Japan we are able to provide professional contact to our more than 200+ clients in the pharma and biotech industry. CGX will provide information throughout the following channels:

- Collaborative Partners: CGX' distinguished partners, research collaborators and contracted partners (e.g. Bayer, Aventis and Merck)
- Client Network: CGX' well established and maintained communication routes to its customers throughout major sales and marketing contacts.
- Conferences and Trade Shows
- Web Page / Affiliated companies: Web presence of the OpenMolGRID project in various locations of CGX' and its sister companies' web sites.

Exploitation methodology:

- Collaborative links to other projects: The collaboration will be for the calculation of chemical descriptors using the OpenMolGRID tools. This way, performances of the OpenMolGRID tools will be checked and compared with results obtained independently within the other projects.
- Collaborative Partners: CGX will introduce the system to its collaborative partners through mutual visits, presentations and discussions and ensure that collaborative partners are kept up to date with developments in the program. Their feedback will be included in the periodic reports.
- Client Visits: CGX and its representatives routinely visit clients. During these visits clients will receive introductory information related to the system and for those who have expressed interest a demonstration will be provided. Upon request the client can receive evaluation access to the system and training materials. Client visit feedback will be included within the periodic reports.
- Client Network: CGX will inform its client base and follow-up interest by providing evaluation access to the system and training materials. Their feedback will be included in the periodic reports. The particular approaches of the dissemination of the news and information regarding the OpenMolGRID system/project to CGX' clients will include:
    - Organizing net meetings with CGX' partners. CGX' scientific professionals will present updated information and results throughout computer and telephone conference aided slide shows. During the presentation, the most attractive topics will be discussed, subject to their interests. The Client Feedback Form will be used as a follow-up to the meeting.
    - Incorporation of the OpenMolGRID Project Summary and Client Feedback Form into CGX' official *Scientific Newsletter*, which is a representative publication informing CGX' clients about the latest scientific news and innovations. The Newsletter is usually printed 1000 number of copies and released by the four seasons.

  - Incorporation of the OpenMolGRID Project Summary and Client Feedback Form into CGX' other *scientific publications* where appropriate.
  - Incorporation of the OpenMolGRID Proje ct Summary and Client Feedback Form into CGX' sales and marketing *CD* as "CGX' Latest Scientific News". The CD comprises sales and marketing information including updated information on the CGX' Compound Repository Stock, animated company information, the latest Press Releases, and scientific updates including the latest Scientific Newsletter, scientific publication request form as well as technical information. The CD's are mailed to most of our clients systematically every month.

- Conferences and Trade Shows: CGX will continuously present the latest news and information on the OpenMolGRID of relevance to those who would use the system. CGX will include scientific details of the development of the system in relevant scientific presentations and posters. Interested attendees can receive evaluation access to the system and training materials. Conference feedback will be included within the periodic report.
- Web Page: Project updates throughout the latest OpenMolGRID Project Summary and links for additional information will be published at CGX' and its affiliated companies' web sites. An easy-to-use environment will provide the possibility to the clients to fill out and submit the Client Feedback Form and comment the development. In the later stage of the project work plan, users can also request evaluation access to the system and training materials. Visitors' feedback will be recorded and included in the periodic reports.

Exploitation Schedule:

The process of the exploitation of results and updated information regarding the OpenMolGRID system/project consists of three stages, as follows.

- The first stage upon completion of the specification stage of the project (T1.1, T2.1, T3.1, T3.2.1, T4.1.1, T4.1.2, T4.2.1).
- The second stage commences with the completion of *functional* testing of the system (T5.1A).
- The final stage commences with the completion of *in silico* testing of the system (T5.2) and the generation of the evaluation version of the system.

In accordance with the requisites for application, CGX will expose evaluation versions of the system to the existing exploitation channels, as outlined above, and deliver feedback in the form of a periodic report scheduled (deliverables D6.4, D6.5, and D6.6).

9.8. WP7: Project Management and Coordination

The project management of OpenMolGRID will use state of the art tools to ensure timely delivery of results. Project tracking will be performed using MS Project. The communication amongst the project partners will use a proven electronic system for collaborative work (Basic Support for Cooperative Work - BSCW; http://www.orbiteam.de/) installed at the University of Tartu. The project management relies on an experienced administration with a proven track record in managing large research and development projects (national and international) and an equivalent financial department.

Management and Coordination

The main objectives of the OpenMolGRID project management are:

- Initiate a goal-directed development environment;
- Facilitate cooperation between partners;
- Raise awareness of quality;
- Detect areas of potential problems early;

- Anticipate and manage change.

To achieve these objectives, the procedures discussed below will be established.

Decision Structures

A Steering Committee consisting of one representative from each principal contractor will be formed. The project will run under the control of the Project Coordinator, the Project Technical Coordinator, and the Project Steering Committee. They are supported by a Quality Engineer.

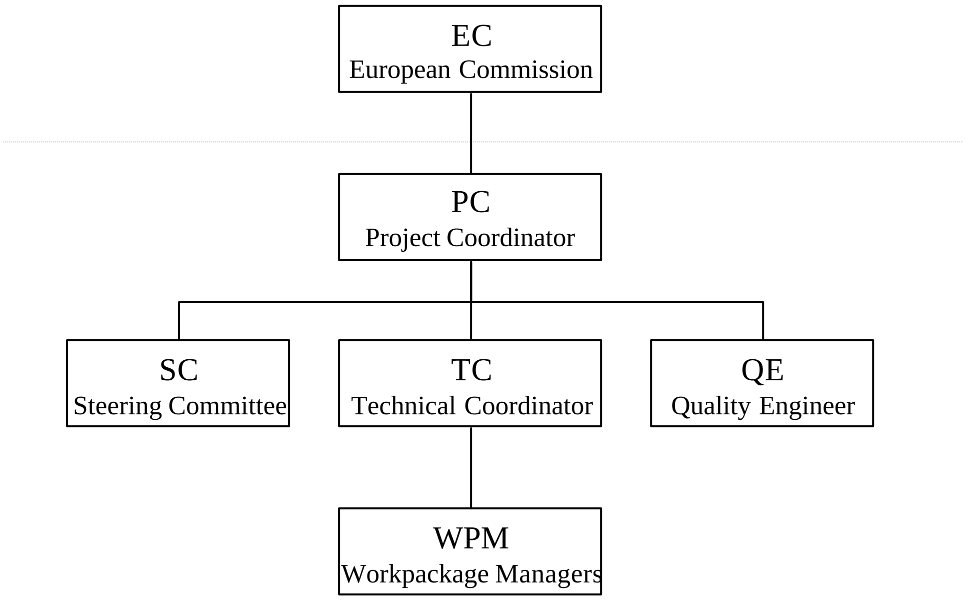

Figure 1: Decision Structure

Communication Flow

Most of the communication within the project will be handled through e-mail with a repository for project documents being established by the coordinator. Project meetings are planned every three month.

Exchange of contractual and financial material is by written documents only.

The project management will encourage frequent and informal communication between the partners in order to facilitate the collaboration, and to quickly devise solutions for technical or support questions.

Project Control

The following measures will be taken:

- Formal reporting of progress, anticipated problems and deviations from plan by the partner responsible for each workpackage every three month;
- Formal report of expended effort for each task by all partners every three months;
- Production of a progress report for the Commission every three month, consolidating the three-monthly progress data;
- Production of a management report for the Commission every three months, consolidating the three-month effort data;
- Prompt reaction to any significant deviation from plan: project board decision on how to proceed, information of the Commission;

Quality Assurance Measures

A Quality Engineer will be appointed who will define standard for all reports and deliverables afte r the project start. Prior to delivery to the Commission, all deliverables will be reviewed internally by at least the responsible Workpackage Manager, the Technical Coordinator, and the Quality Engineer. The project management will make sure that this process is started in time, and that the finished deliverables are transferred to the Commission in time.

For deliverables that contain software prototypes, a special procedure will be followed that makes sure that the respective prototypes can be installed and used by all concerned partners.

External Reviews

It is anticipated that external reviews will be held every ten month. The project management will make sure that all local preparations are made, and send out invitations at least 28 days in advance. Deliverables due at a review will be transferred to the Commission at least 14 days in advance.

10. List of Relevant Publications

General publications relevant to OpenMolGRID:

1. Foster, I., Kesselman, C., "The Grid: Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, Inc., San Francisco, 1999.
2. Weiss, G., (ed), Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, The MIT Press, Cambridge, MA; London, England, 1999.
3. JSR-73 Expert Group, "Java(TM) Specification Request 73: Java Data Mining (JDM)", Community Review Draft 0.80 (W. Dubitzky is member of the JSR-73 Expert Group; the document is currently not publically available).
4. Object Management Group, "The Common Warehouse Metamodel (CWM(TM))", http://www.omg.org/technology/cwm/
5. The Data Mining Group, Predictive Model Markup Language (PMML), http://www.dmg.org/
6. R. G. G. Cattell, Douglas K. Barry, Mark Berler, Jeff Eastman, David Jordan, Craig Russell, Olaf Schadow, Torsten Stanienda, and Fernando Velez, "The Object Data Standard", Object Database Management Group (ODMG) 3.0, Morgan Kaufmann Pub., 2000.
7. Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite, "The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses", Wiley, 1998.
8. B. Allcock, J. Bester, J. Bresnahan, A. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnel, S. Tuecke, "Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing", in Proceedings of IEEE Mass Storage Conference, 2001.
9. Grimshaw, Wm.A. Wulf, and the Legion team, "The Legion Vision of a Worldwide Virtual Computer", in Communications of the ACM, Vol.40, No.1, pp39-45, January 1997.
10. M. Thompson, W. Johnston, S. Mudumbai, G. Hoo, K. Jackson, A. Essiari, "Certificate based access control for widely distributed resources", in Proceedings of the Eighth Usenix Security Symposium, 1999.
11. R. Buyya, "Economic-based Distributed Resource Management and Scheduling for Grid Computing", PhD thesis, School of Computer Science and Software Engineering, Monash University, Melborne, Australia, April 2002.
12. R. Buyya, K. Branson, J. Giddy, D. Abramson, "The Virtual Laboratory: Enabling Molecular Modelling for Drug Design on the World Wide Grid Technical Report", Monash-CSSE-2001-103, Monash University, December 2001.
13. G. Allen, W. Benger, T. Dramlitsch, T. Goodale, H.C. Hege, G. Lanfermann, A. Merzky, T. Radke, E. Seidel, "Cactus Grid Computing: Review of Current Development" in Proceedings of 7th International Euro-Par Conference Manchester, LNCS 2150, p. 817 ff, UK, August 2001.

Relevant publications from University of Tartu:

1. S. Sild, M. Karelson, "A General QSPR Treatment for Dielectric Constants of Organic Compounds", J. Chem. Inf. Comput. Sci., 42(2), pp360-367, 2002.
2. Lomaka, M. Karelson, "A pivot algorithm for generating lowest energy structures of peptides", Chemical Physics Letters, Volume 346, Issues 3-4, pp 322-328, 2001.

3. Karelson M, Molecular Descriptors in QSAR/QSPR, J. Wiley & Sons, New York, 2000.
4. M. Karelson, S. Sild, U. Maran, "Non-linear QSAR Treatment of Genotoxicity", Molecular Simulations, 24, pp229-242, 2000.
5. M. Karelson, U. Maran, Y. Wang, A.R. Katritzky," QSPR and QSAR Models Derived with CODESSA Multipurpose Statistical Analysis Software", AAAI Tech. Report, SS-99-01, 12-23 (1999).
6. M.C. Menziani, M. Montorsi, P.G. De Benedetti, M. Karelson, "Relevance of Theoretical Descriptors in QSAR Analysis of G-protein Receptor Antagonists", Bioorg. & Med. Chem., 7, pp2437-2451, 1999.
7. M. Karelson, G.H.F. Diercksen, "Models for Simulating Molecular Properties in Condensed Systems", in "Problem Solving in Computational Molecular Science: Molecules in Different Environments", S. Wilson and G.H.F. Diercksen (Eds.), Kluwer Academic Publ., Dordrecht, 1997, 215-248.

Relevant publications from University of Ulster:

1. W. Dubitzky, A. Schuster, D.A. Bell, J.G. Hughes, K. Adamson, "How Similar is VERY YOUNG to 43 Years of Age? On the Representation and Comparison of Polymorphic Properties", in Proc. 15th Int'l Joint Conference on Artificial Intelligence, pp226-231, Japan, August 1997.
2. W. Dubitzky, F. Azuaje, "Chapter 6: A Soft Computing Approach to Modelling and Learning Case Retrieval Structures", in Soft Computing and Case-Based Reasoning, Sankar K. Pal (ed.), Tharam S. Dillon (ed.), Daniel S. Yeung (ed.) Springer-Verlag, pp115-146, 2000.
3. W. Dubitzky, M. Granzow, D. Berrar, "Comparing Symbolic and Subsymbolic Machine Learning Approaches to Classification of Cancer and Gene Identification", in Simon M. Lin and Kimberly F. Johnson (editors), Methods of Microarray Data Analysis: Papers from CAMDA'00, pp151-166, Kluwer Academic Publishers; ISBN: 0792375645, 2001.
4. W. Dubitzky, M. Granzow, D. Berrar, "Data Mining and Machine Learning Methods for Microarray Analysis", in Simon M. Lin and Kimberly F. Johnson (editors), Methods of Microarray Data Analysis: Papers from CAMDA'00, pp5-22, Kluwer Academic Publishers; ISBN: 0792375645, 2001.
5. Berrar D., Dubitzky W., Solinas-Toldo S., Bulashevska S., Granzow M., Conrad C., Kalla J., Lichter P., Eils R., "Design and Implementation of a Database System for Comparative Genomic Hybridization Analysis", in IEEE Engineering in Medicine and Biology, Vol. 20, Number 4, pp75-83, July/August, 2001.
6. F. Azuaje, F., W. Dubitzky, N. Black, K. Adamson, "Improving Clinical Decision Support Through Case-Based Fusion", in IEEE Transactions on Biomedical Engineering, Special Issue on Biomedical Data Fusion, 46 (10) 1181-1185, 1999.
7. Azuaje, F., W. Dubitzky, N. Black, K. Adamson, "Discovering Relevance Knowledge in Data: A Growing Cell Structures Approach", in IEEE Transactions On Systems, Man And Cybernetics. Part B: Cybernetics, Vol. 30, No 3, pp448-460, 2000.

Relevant publications from Mario Negri Institute:

1. G. Gini, V. Testaguzza, E. Benfenati, R. Todeschini, "HyTEx (Hybrid Toxicology Expert system): architecture and implementation of a multi-domain hybrid expert system for toxicology", Chemometrics and Intelligent Laboratory Systems, 43, pp135-145, 1998.
2. Giuseppina Gini, Marco Lorenzini, Emilio Benfenati, Paola Grasso, Maurizio Bruschi, "Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network", J. Chem. Inf. Comp. Sci., 39, pp1076-1080, 1999.
3. G. Gini, E. Benfenati, D. Boley. Clustering and Classification Techniques to Assess Aquatic Toxicity. Procs the 4th Int'l Conf. KES 2000, Brighton, UK, 1, pp166-172, 2000.
4. G. Gini, M. Lorenzini, E. Benfenati, R. Brambilla, L. Malvé, Mixing a Symbolic and a Subsymbolic Expert to Improve Carcinogenicity Prediction of Aromatic Compounds, in: Multiple Classifier Systems, Ed: J. Kittler, F. Roli, Springler-Verlag, Berlin, pp126-135, 2001.
5. A.R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, E. Benfenati, M. Karelson, U. Maran, "Interpretation of quantitative structure-property and -activity relationships", in Journal of Chemical Information and Computer Sciences, Volume 41, Issue 3, pp679-685, 2001.
6. G. Sello, L. Sala, E. Benfenati, "Predicting toxicity: a mechanism of action model of chemical mutagenicity", in Mutation Res., 479, pp141-171, 2001.
7. E. Benfenati, N. Piclin, A. Roncaglioni and M. R. Varì, "Factors Influencing Predictive Models For Toxicology", in SAR and QSAR in environmental research, 12, pp593-603, 2001.

Relevant publications from Forschungszentrum Jülich:

1. M. Romberg, "The UNICORE Grid Infrastructure", To appear in Scientific Programming Vol.9 No.4, ISSN 1058-9244, 2002.
2. W.E. Dietmar, D.F. Snelling, "UNICORE: A Grid Computing Environment", in Proceedings of 7th International Euro-Par Conference Manchester, LNCS 2150, p825 ff, UK, August 2001.
3. V. Huber, "Supporting Car-Parrinello Molecular Dynamics with UNICORE", in Proceedings of International Conference on Computational Science – ICCS 2001, Pt. 1, Springer Verlag, Vol. 2073, pp560-567, May 2001.
4. M. Romberg, "The UNICORE Architecture: Seamless Access to Distributed Resources", Proceedings of the eighth IEEE International Symposium on High Performance Distributed Computing, ISBN 0-7803-5681-0, pp287-293, August 1999.
5. V. Sander, D. Erwin, V. Huber, "High-Performance Computer Management Based on Java", in Proceedings of the HPCN '98, 21.4. - 23.4.1998, Amsterdam, pp526–534, 1998.
6. J. Grotendorst, D. Marx, A. Muramatsu (Hrsg.), "Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms - Lecture Notes", Winterschule, 25. Februar - 1. März 2002, Kerkrade, ISBN 3-00-009057-6, 548 pages, February 2002.
7. R. Esser, P. Grassberger, J. Grotendorst, M. Lewerenz, (eds.), "Proceedings of the Workshop on Molecular Dynamics On Parallel Computers", Research Centre Jülich, 8 - 10 February 1999, World Scientific 2000, ISBN 981-02-4232-8, 379 Pages.

Relevant publications from ComGenex, Inc:

1. Á. Papp, T. Fujita, F. Darvas, "Design of Focussed Libraries Using "Bioanalogous" Transformation Rules Implemented to an Expert System (EMIL)", in Eurocombi-1 Symposium, Budapest, Hungary, July 1-5, 2001.
2. F. Darvas, T. Fujita, A. Papp, "Introduction of a Web-based Tool to Build Bioanalogous Libraries for Lead Optimization", in CHI Drug Discovery Japan, Tokyo, Japan, Jan. 26-Feb.2 2002
3. F. Darvas, I. Szabó, Gy. Dormán, "High-Throughput Combinatorial Chemistry Combined with Predictive Tools: Application in Early Metabolism/Retrometabolism Studies", in 6th European Congress of Pharmaceutical Sciences, Budapest, Hungary, September 16-19, 2000.
4. F. Darvas, "Optimal Integration of Predictive and In Vitro Experimental Approaches in Early ADME/Tox Prediction", in Early High Throughput ADME & Toxicology Studies Conference, Basel, Switzerland, March 8th, 2000.
5. F. Darvas, Gy. Dormán, "Early Integration of ADME/Tox parameters into the design process of combinatorial libraries", in Chimica Oggi/Chemistry Today, pp10-13, July/August 1999.
6. F. Darvas, S. Marokházi, P. Kormos, G. Kulkarni, H. Kalász, Á. Papp, "Metabolexpert: Its Use in Metabolism Research and in Combinatorial Chemistry", in Drug Metabolism, Databases and High-Throughput Testing During Drug Design and Development, pp237-270, Ed. by P.W. Erhardt, IUPAC, Toledo, Ohio, 1999.
7. F. Darvas, Gy. Dormán, Á. Papp, "Diversity Measures for Enhancing ADME Admissibility of Combinatorial Libraries", in J. Chem. Inf. Comp. Sci., Vol. 40, No. 3, pp314-322, 2000.

Relevant publications from Geerd HF Diercksen:

1. S. Yamamoto, H. Tatewaki, O. Kitao, G. H. F. Diercksen, 'Rydberg character of the higher excited states of free base porphin", in Theor. Chem. Acc., 106, pp287-296, 2001.
2. W. Duch, R. Adamczak, G. H. F. Diercksen, "Constructive density estimation network based on several different separable transfer functions", in: 9th European Symposium on Artificial Neural Networks, Brugge 2001, De-facto Publications, pp107-112, 2001.
3. W. Duch, R. Adamczak, G.H.F. Diercksen, "Classification, Association and Pattern Completion using Neural Similarity Based Methods", in Applied Mathematics and Computer Science, 10, pp101-120, 2000.
4. S. Yamamoto, M. Karelson, G. H. F. Diercksen, "An ab-initio CI Study of the Electronic Spectra of Substituted Free-base Porphins", Chem. Phys. Letters, 318, pp590-596, 2000.
5. O.N. Ventura, M. Kieninger, S. Suhai, G. H. F. Diercksen, "The Water Dimer: Post-Hartree-Fock And Density Functional Calculations on the Potential Energy Surface", in: Molecular Engineering, 7, pp317-348, 1997.
6. M. Karelson, G. H. F Diercksen, "Models for simulating molecular properties in condensed systems", in: Problem Solving in Computational Molecular Science: Molecules in Different Environments, S. Wilson and G. H. F. Diercksen (Editors.), NATO ASI Series, Series C: Mathematical and Physical Sciences - Vol 500, Kluwer Academic Publishers., Dordrecht, p215-248, 1997.
7. G. H. F. Diercksen and G. G. Hall, "Intelligent software: The OpenMol program", in: Computers in Physics, 8, 215-222 (1994); Lecture Notes in Computer Science 796, W. Gentzsch and U. Harms (Eds.), Springer-Verlag, Berlin, p 219-222, 1994.