CHIMIA **2004**, *58*, No. 9

Chimia 58 (2004) 577–584

ISSN 0009–4293

Molecular Informatics as an Enabling *in silico* Technology Platform for Drug Discovery

Edgar Jacoby*, Ansgar Schuffenhauer, Maxim Popov, Kamal Azzaoui, Eric Vangrevelinghe, John Priestle, Philippe Ferrara, Bernard Faller, and Pierre Acklin

Abstract: The molecular informatics platform, as implemented today in the Molecular and Library Informatics (MLI) Technology Program at Novartis Institutes for BioMedical Research (NIBR) Discovery Technologies, will be presented. The mission of the MLI program is primarily defined to contribute to the selection of screening hit and lead compounds using *in silico* methods. The MLI technology program aims to provide an integrated pipeline of computational methods for high-throughput *in silico* screening combining specific cheminformatics, bioinformatics, docking and 3D pharmacophore applications. The four core activities of the group include: 1) Molecular diversity management; 2*) In silico* screening using HTD (high-throughput docking) and 3D pharmacophore searching; 3) Integrated analysis of HTS (high-throughput screening) and profiling data; and 4) Database management and software engineering in the field of *in silico* screening. The contribution of these activities to the drug discovery process will be summarized together with novel trends in the field.

Keywords: Data analysis · *in silico* Screening · Molecular diversity · Molecular informatics

1. Introduction

Molecular informatics platforms are currently emerging as a new enabling technology which aims to integrate key cheminformatics and bioinformatics processes for drug discovery. Molecular informatics can be defined as the integration of the biological and the chemical knowledge spaces and is recognized as a foundation for the improvement for the quality of informatics and high-throughput sciences driven drug discovery [1][2]. The establishment of standardized molecular informatics platforms is thus consequently pursued within the academic and industrial drug discovery organizations and informatics-based discovery technology companies [3–5]. The Novartis Institutes for Biomedical Research, Molecular and Library Informatics (NIBR MLI) platform described here, focuses on lead discovery, which is an essential element of the industrial drug discovery process. Molecular diversity management aims to enrich and enhance the diversity and focus of the NIBR screening collection by the selection of new lead-like compounds for worldwide compound purchasing and library design campaigns. In the perspective of chemogenomics, which aims to identify systematically all ligands and modulators for all expressed gene products, the vision is to match the chemistry and biology spaces [6]. In addition, this activity aims for a compilation of target-family focused and diverse screening sublibraries from the comprehensive NIBR screening collection for classical and fragment-based screening approaches. High-throughput docking (HTD) and 3D pharmacophore searches are focused on the selection of screening compounds based on the 3D structural hypotheses of ligand–receptor interactions provided by structural biology. Both virtual screening methods are used in an independent and in an integrated manner within the high-throughput screening (HTS) process flow [7][8]. Integrated analysis of experimental HTS and profiling data focuses on the analysis of the chemical and biological information content of quality control (QC) validated HTS and profiling data. By providing tools for filtering, clustering and similarity searching of comprehensive bioprint-like structure–activity tables in both the chemical and biological dimensions, we enable the rational selection of hit and lead compounds. The database management and software engineering activity develops the necessary informatics tools in collaboration with internal and external collaborations; an outline for software developments required in the future will be provided.

*Correspondence: Dr. E. Jacoby
Novartis Institutes for BioMedical Research
Discovery Technologies, Compound Logistics and Properties Unit
CH–4002 Basel
Tel.: +41 61 32 46186
Fax: +41 61 32 46261
E-Mail: edgar.jacoby@pharma.novartis.com.

2. Molecular Diversity Management

The NIBR compound collection enrichment and enhancement project integrates corporate internal combinatorial compound synthesis and external compound acquisition projects in order to build up a comprehensive screening collection for a modern global industrial drug discovery organization. The main purpose of the screening collection is to supply the Novartis drug discovery pipeline with hit-to-lead compounds for today's and the future's portfolio of drug discovery programs and to provide tool compounds for the chemogenomics investigation of novel biological pathways and circuits. As such, it integrates designed focused and diversity-based compound sets from the synthetic and natural paradigms able to cope with druggable and what are currently deemed to be undruggable targets and molecular interaction modes [9].

Molecular informatics is a key component for the efficient management of the compound collection enhancement and target family focusing activities. The assessment of the likelihood of a molecule to bind to a molecular target is equally important for both activities. Compared to protein structure-based approaches like HTD, ligand-based similarity and diversity approaches can be applied routinely and in a rapid manner to the quite large physically existing and virtually designed compound sets typically available for selection campaigns [8][10]. Today, most of the existing similarity searching methods are used to identify candidate screening compounds for a target where reference compounds are already known – allowing competitors to find catch-up lead molecules. Cheminformatics similarity searching methods able to identify not only ligands binding to the same target as the reference ligand(s), but also potential ligands of other homologous targets for which no ligands are yet known, are essential tools for the further exploration of the previously successful target families. Such methods have emerged only recently and most of them are based on self-organizing maps created on different types of molecular descriptors [11–13].

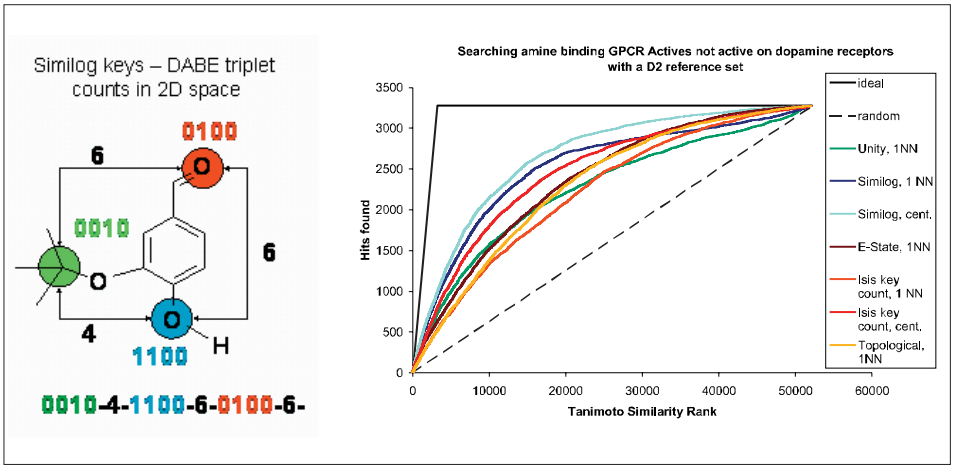

Fig. 1. The Similog keys and retrospective *in silico* screening experiments comparing the retrieval performance of this molecular descriptor with others for chemogenomics homology-based similarity searching applications. The Similog keys are counts of atom triplets where each triplet is characterized by the 2D interatom graph distances and the types of its atoms. The atom-typing scheme classifies each atom by its function as H-bond donor (D) or acceptor (A), and by its bulkiness (B) and electropositivity (E). Other descriptors used for the validation included: the Unity-2D fingerprints, 2D topological descriptors, ISIS public key count and the E-state descriptors. These descriptors were used in combination with two Tanimoto distance averaging methods: 1NN – nearest neighbor and cent. – centroid. The MDDR data set was split randomly into two halves of which the first was used to obtain reference sets and the second was used as test set. Within the illustrated enrichment curves, D_2 GPCR ligands were used to search for ligands of the other monoamine GPCRs excluding all dopamine GPCR hits. For further detail see [14].

Within NIBR, we have designed a method called homology-based similarity searching which is based on the Similog molecular descriptor developed previously at Sandoz. The method consists of the following three steps: 1) Select at least one target with known ligands that is homologous to the new target; 2) combine the known ligands of the selected target(s) to a reference set; 3) search candidate ligands for the new targets by their similarity to the reference set [14]. Our approach has been validated using retrospective *in silico* screening experiments on datasets of the MDL Drug Data Report (MDDR) catalogue for several target families, including the monoamine G-Protein Coupled Receptors (GPCRs) as illustrated in Fig. 1.

Although we observed in our study fewer differences among the descriptors for similarity searching towards the same target as the reference target, the application of the Similog keys is more effective in the identification of ligands binding to targets homologous to the reference target. We attribute this superiority to the fact that the Similog keys provide a generalization of the chemical elements and that the keys are counted instead of merely noting their presence or absence in a binary form. The Similog keys are thus capturing the potential points of conserved interactions between the ligands and the target proteins. The difference in the performance of the distance averaging methods is attributed to the fact that especially the centroid method is able to enhance commonalities displayed in the pharmacophore representations of the reference compounds, crystallizing in this manner the main repeated pharmacophore features. The development and evaluation of novel chemical descriptors for *in silico* screening applications is a highly rewarding field of cheminformatics research and is consequently pursued with high priority in our group and collaborations [15][16].

In the process of building a focused library, it has turned out to be advisable to carefully review the list of the reference compounds extracted from the database together with the project chemists in order to remove reference compounds with unwanted mechanisms such as covalent binders or frequent and promiscuous hitters, and in order to add relevant reference compounds from the corporate history [17–19]. The similarity searching is post-processed to avoid having too many compounds brought by similarity to one individual reference compound and clustering is used to ensure the maximum possible diversity of chemotypes in the resulting focused screening set. In accordance to the findings of others, our experience with focused screening sets for kinase, proteases and GPCRs is that the results are very positive and that hit rates of 1–10% covering multiple chemical chemotypes can be expected with library sizes of 500–2500 compounds, when the libraries are designed towards new members with expected conserved molecular recognition [20][21]. Based on our experience, the screening of focused libraries is invaluable, both very early in the discovery process in order to generate tool compounds for target validation, and also in combination with HTS, which, based on the statistical nature of the procedure, cannot necessarily be expected to identify every active compound. Consequently, a number of target family focused sets are maintained and implemented for new emerging target families; the concept is used to structure the NIBR screening collection.

True molecular information systems capturing up-to-date knowledge of ligand and target data are essential for the compilation of comprehensive reference compounds sets. The classical ligand database systems like MDDR, Ensemble, WDI (World Drug Index), CMC (Comprehensive Medicinal Chemistry), IDdb (Investigational Drugs database) or PharmaProjects provide structural information of ligands together with therapeutic class information and sometimes molecular target information. The latter information is however provided without any further phylogenetic or other relationship among the targets, which limits their value for chemogenomics applications. Given these limitations, we implemented a ligand–target ontology for ligands of four major target families of interest to us. The system enabled us to collate systematically ligands to comprehensive reference sets for any specified levels of classification.

The MDDR database constitutes the underlying ligand dataset and the ligand–tar-

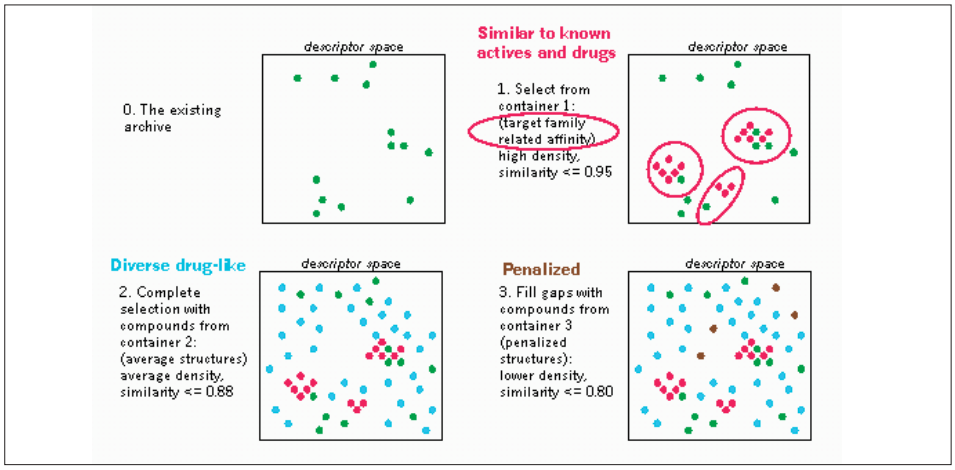

Fig. 2. Schematic illustration of the NIBR compound selection process. The candidate compounds are in a first step filtered using substructure and computed physicochemical filters. Unwanted, highly reactive and toxic compounds like carboxylic acid anhydrides are eliminated. The remaining compounds are divided iteratively into three orthogonal categories. Penalized compounds are defined based on additional substructure filters identifying particularly abundant chemotypes like 4-phenyldihydropiperidines. Using both Unity 2D fingerprints and Similog keys, the remaining compounds are compared to selected known reference drugs and actives of the main target families of interest to NIBR and separated into the category of similar to known drugs and the category of drug-like diverse compounds. The candidate compound sets are then compared in an incremental manner to the existing collections and the previously made selections. This diversity selection process starts with the candidate set of similar to known drugs and ends with the penalized compound set, and is done using the Tripos Optisim algorithm [30]. The incremental diversity selections are done with deceasing similarity thresholds and compound densities. For further detail see [28].

get annotation is based on the classification references established by the EC (Enzyme Commission), GPCRDB (G-Protein Coupled Receptor Database), NuclearDB (Nuclear Receptor Database), and LIGCDB (Ligand Gated Ion Channel database). By linking MDDR activity keys to the targets of the classification schemes, we were able to assign and group 50'000 of the 100'000 MDDR ligands with their macromolecular target and target classes, in addition to the ligand functional class and therapeutic indication [22]. During this activity we recognized that one main unaddressed challenge is to provide such annotations at the protein binding site and genome wide levels [23][24]. It is noteworthy that chemogenomics knowledge-based companies like Aureus Pharma, Inpharmatica, GVK-Biosys, Sertanty, and Jubilant are establishing comprehensive molecular information systems for a variety of target classes, including GPCRs, kinases, ion channels, and proteases. In addition, the value of annotated compound libraries was recently emphasized as a chemical tool kit for the investigation of novel disease-relevant signaling pathways and cellular circuits [25].

Similar selection methods can be used to assess drug-similarity for the general enhancement of the screening collection by compound purchase selections. The structural diversity is of particular importance here, not only exact duplication needs to be avoided, but a general diversity in terms of chemical classes, lead and drug-likeness needs to achieved [17][26]. The NIBR screening library contains several types of subsets, including annotated known bioactive compounds; target-family focused libraries (*e.g.* kinases, proteases, *etc.*); peptide mimetics (*e.g.* β-strand, β-turn, α-helix structural mimetics); natural products and derivatives thereof, and diversity-oriented synthesis (DOS)-based libraries which tend to mimic the structural complexity and the skeletal and stereochemical diversity of natural products [27].

The NIBR selection process is informatics-, chemistry- and biology driven and consists of two steps (Fig. 2) [28]. In the first step, the candidate compounds are filtered and grouped into three priority classes on the basis of their individual structural and computed physicochemical properties. Substructure and computed physicochemical filters are used both to eliminate and to penalize compound classes. The similarity of the remaining structures to selected reference ligands of proven druggable target families of interest is then computed, and the compounds similar to drugs and known actives are prioritized for the following diversity analysis; homology-based similarity searching is thus used here with several reference sets of the major target families of interest to the current target portfolio simultaneously. In the second step, the compounds are compared to the archive compounds and a diversity analysis is performed. This is done separately for the compounds prioritized as similar to known drugs and actives, the drug-like regular compounds and the penalized compounds with increasingly stringent dissimilarity criteria. The automated analysis is followed by manual review of the compounds to assess more complex structural properties like the chemical derivatization potential; one major role of cheminformatics is thus recognized in the need to reduce the number of potential candidates to a humanly tractable number [29].

Regarding the assessment of chemical diversity, recent advances in clustering techniques are noteworthy [31–33]. They now enable the co-clustering of very large commercial compound collections and reference sets with the entire corporate collection and allow the application of constraints for the minimal number of compounds to be selected per cluster. The ideal library size is currently a subject of scientific debate. Whereas theoretical rationales are emerging [34–36], pragmatic considerations are prevailing and focus on the diversity of chemotypes rather than on larger and larger numbers of individual compounds per scaffold; the latter should, however, be such to enable the detection of structure–activity relationship from the screening data. As an increasing number of commercially available screening compounds are prepared by combinatorial or parallel synthesis, the evaluation and selection based on the scaffolds is a reasonable alternative. This is especially valid if the compounds have not yet been prepared and one is given the opportunity to prioritize the synthesis proposals. Scaffold novelty within the corporate collection and compared to the patented chemistry space can be ensured by substructure searching. As the number of attractive scaffolds is limited, the selection of the most promising ones can be done manually, although computational methods for the evaluation of scaffold diversity are emerging [37–38]. Reference repertoires of privileged structures are a pragmatic guide in this process [39].

The implementation of efficient and updated 2D- and 3D-structure databases is one major challenge in molecular diversity management. Although dedicated cheminformatics companies are providing updated unified compilations of the major vendor catalogues [40], corporate internal databases are needed, both for compounds and scaffold-based projects because many specialized vendors only share information on a confidential basis. Databases of around 5'000'000 compounds and around 4'000 scaffold-based chemistries available inside NIBR as well as from selected vendor catalogues were thus created with the main purposes to facilitate the selection and purchase logistics of new compounds and to assist chemists to follow-up screening hits. Each compound in the 2D database is rep-

resented with its unique structure and sample ID, which can be a vendor catalogue number or a Novartis internal number. For each entry, the original structure representation is saved and the structural representation is standardized after stripping out all solvent and salt fragments using data pipelining tools [41]. The information about solvent or salt fragments is stored together with the sample ID in a coded form, and the standardized structure is entered with the link to the corresponding sample. The representation of samples as standardized structures allows quick identification of structural duplicates between purchase candidates. A number of structural properties and fingerprints are calculated and stored to monitor the properties of the screening collection and to enable similarity and diversity analysis for compound selection.

As the current structure database technology is reaching its performance limits for index updates and database searching, new developments are required especially if virtual compound sets become available which are potentially several order of magnitudes larger than the physically available sets. Such new developments should ideally include search capabilities based on multiple complementary molecular descriptors [10].

3. *In silico* Screening Using HTD and 3D Pharmacophore Searching

Given the availability of Linux clusters and Grid computing platforms, HTD and 3D pharmacophore *in silico* screening techniques have matured during the last years to become reliable, inexpensive and fast methods for lead finding which complement HTS [42–45]. HTD is applicable when a relevant 3D model of the target structure is available. The 3D query for pharmacophore searching can be based on both the 3D target and ligand information. Both methods try to optimize and rank the complementarity between the candidate compound and the 3D structure of the binding site or the pharmacophore query. Only those compounds with the highest complementarity are then actually biologically tested. The final selection of compounds often integrates consensus scoring techniques [46][47] and careful visual inspection by a computational and medicinal chemist. A variety of docking and pharmacophore searching programs are commercially available, the most prominent being Glide [48][49], Gold [50], FlexX [51], ICM [52], Catalyst [53], and Unity [54].

Within the MLI program, HTD and pharmacophore searching are applied for different scenarios and over the last two years, resulted in a significant number of hit-to-lead compounds. The methods are used to screen compound catalogues from medicinal chemistry vendors, enriching the compound acquisition campaign by compounds with an increased probability to bind to specific targets of interest. Both for classical screening and fragment-based screening methods, specific target screening boxes are generated and subsequently screened, providing additional hit series compared to HTS. In many drug discovery projects, a 3D structure becomes available only after the HTS is completed. HTD is of value to probe the newly available chemistry space when a new follow-up lead series is required and a large-scale HTS campaign is not a priority. For instance for an important kinase target, using HTD to screen compound libraries recently added to our archive, we were able to discover 16 new validated inhibitors with different chemotypes from only around 200 compounds selected for testing. HTD and pharmacophore-based *in silico* screening are also conducted in a more 'classical' way in exploratory projects or when a HTS is not available, for instance for complex cell-based assays for virus fusion. We identified in a early phase of the CK2 program a novel ATP binding site inhibitor using a homology model of the target [55]. The complementarity of HTS and HTD is of particular importance [8][56], especially in the factory HTS set-up of NIBR Discovery Technologies. HTD is an essential element of the HTS data analysis process in order to rescue promising potential false negatives that are below the traditional HTS threshold, but which show steric and electrostatic complementary to the postulated target binding site. Applying this approach in a recent HTS campaign, around 100 validated actives and 120 weak actives out of 500 rescued compounds were found in addition to those validated hits from the initial selection made by screening (Fig. 3). Another use of HTD is the *in silico* screening of diverse and targeted virtual combinatorial libraries. In lead finding, we use HTD for computational positional scanning experiments of given known actives from internal or competitor discovery projects where there is a need to improve the affinity and/or ADME profile. For an important protease target, we have for instance built a virtual library of 10'000 compounds by grafting on one substituent position commercial reagents to a small tightly binding scaffold. Of the final 20 selected compounds for synthesis, four turned out to be active, with the most active resulting in a lead compound. Such results clearly demonstrate that HTD has become a reliable lead discovery technology.

Given the progress in the integration of structural genomics [5][57], including homology modeling [58][59] and industrial protein structure groups solving quickly the 3D structures of relevant drug discovery targets, the impact of structure-based *in silico* screening can only be expected to grow. Recent developments within the chemogenomics field include the application of HTD for the evaluation of combinatorial libraries against multiple targets and the docking of single compounds against the comprehensive protein structure database [60–64].

The development of better HTD methods is an active field of research in compu-

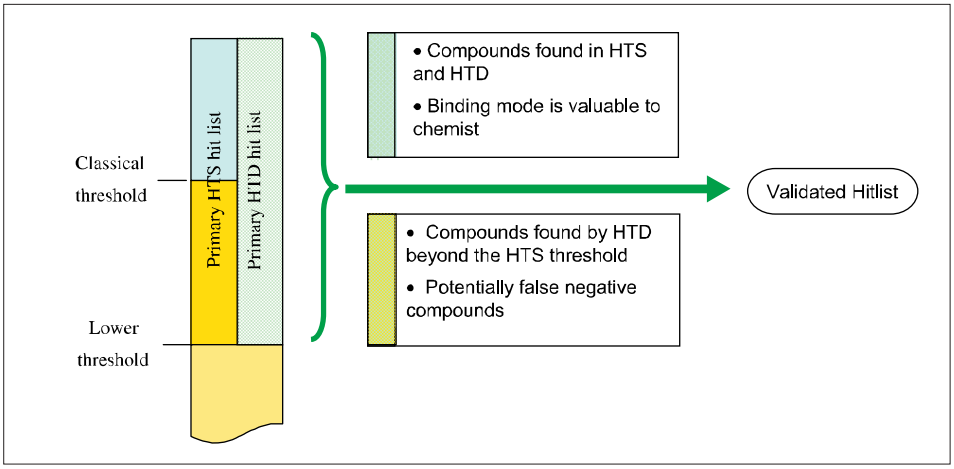

Fig. 3. Process integration of HTS and HTD high-throughput discovery technologies. HTD is used to recover potential false positives which have in the 3D *in silico* model complementarity to the presumed target binding site. The entire screening collection or only compounds from the primary HTS hit list in between the classical HTS thresholds set by the screener and an extended threshold value are investigated in HTD. The additionally selected 500–1000 compounds are added to the validation screening set. Timelines for the generation of the *in silico* input are critical for the success of this process. Additional added value for the drug discovery chemist is provided by the generation of the 3D poses of the validated hits.

tational chemistry [65]. Indeed, the currently available methods are based on a number of approximations regarding the flexibility and the computation of binding energies. It is known that proteins, although depicted as static conformationally averaged representations from crystallographic structure determinations, are actually quite flexible and dynamic molecules. Proteins can undergo small- to large-scale conformational reorganization upon interaction with the ligands as is evidenced from NMR and crystallographic investigations where proteins were crystallized under different conditions or in presence of different ligands [66]. Although docking methods like ICM [52] allow the inclusion of full protein flexibility, at least in the low-throughput mode which is up to a factor 100 times slower, the requirements for high-throughput are such that today the protein is kept fully rigid or that only limited flexibility is allowed for key side-chains [67][68]. Pragmatically, this limitation can partially be circumvented by using multiple protein conformations in parallel generated by NMR structure determinations or MD simulations.

Highly flexible compounds are excluded from large-scale HTD runs because the number of conformations to be generated and checked is computationally too important in a high-throughput mode. Typically, only compounds with twelve or fewer freely rotatable bonds are considered in HTD experiments. To extend the chemical repertoire to natural products is another challenge for HTD, which is important given the particular relevance of natural products for medicine and their limited availability in the screening collection [62][69]. Although natural products are chemicals like synthetic compounds, they tend to be larger and more complex in structure; the number of rotatable bonds tends to be higher and they often contain complex macrocycles. Most docking programs cannot handle the flexibility of such ring systems. In addition, for many natural products the absolute configuration of a number of chiral centers are not known, meaning that the 3D structure cannot be generated with certainty and that a combinatorial enumeration has to be used.

While the docking algorithms have significantly evolved over the years [65], none of the available scoring functions is sufficiently accurate for the reliable computation of binding free energies [46]. The most rigorous computational techniques for binding affinity calculation are the free energy perturbation and thermodynamic integration methods. These methods, which employ molecular dynamics (MD) or Monte Carlo (MC) simulations, are well suited to compute the differences of binding energies of members of a series of congeneric ligands. The drawbacks are that they are computationally very intensive and not practical for ligands that are structurally very different. An approximation to these methods is based on linear response theory (LR) and requires only simulations (MD or MC) at the two endpoints of the binding process, which significantly reduces the computational cost [70]. In this approach, the free energy of binding is given by a weighted sum of the electrostatic and van der Waals interaction energies between the ligand and its environment. A major advantage of the LR approach is that it can handle ligands that are structurally very diverse. Nevertheless, it remains to be seen whether the LR method can yield accurate results in this case. One drawback is that the binding energies have to be scaled by one or more parameters obtained by fitting the computed binding energies to the corresponding experimentally determined binding affinities. Previous work has shown that these fitting parameters depend on the receptor and probably also on the force-field used in the simulation [71].

Another approach that is comparable to LR in terms of computational load is the so-called MMPB/SA (Molecular Mechanics Poisson Boltzmann/Accessible Surface) approach [72]. This method does not require any fitting parameter. Conversely, the solvent is usually modeled as a continuum with a high dielectric value and it remains to be seen whether this approximation is accurate for large biomolecules. As for the LR method, it is unclear to what extent the MMPB/SA can handle ligands that are structurally very different. Very recently, a novel method that combines *ab initio* calculations with suitable consistency restraints has been introduced to compute the ligand partial charges [71]. Used in conjunction with MD simulation in explicit solvent, it has been successfully applied to elucidate the binding mode of progesterone to its receptor [73]. Its value for scoring of HTD hit lists, however, remains to be demonstrated. The methods based on MD or MC simulation are significantly slower than the current scoring functions and can therefore only be applied to the refinement of primary HTD hit lists.

Integrated HTD project management, reporting and database systems which contain both the protein binding site datasets [5] and the ligand datasets will become of interest if the importance of HTD as a discovery technology continues to increase.

4. Analysis of the Chemical and Biological Information Content of HTS and Profiling Data

The global NIBR HTS screening operation generates massive amounts of high-quality screening and profiling data. The primary mission is to enable the selection of high-quality lead structures for the NIBR disease areas by appropriate data analysis. High-quality analysis and visualization of HTS and profiling data are therefore key to successful completion of the mission of NIBR Discovery Technologies and should result in an overall improvement of the quality and speed of discovery transitions and an increase in the research productivity [74][75]. Furthermore, the exploration of the potential of in-depth analysis of the chemical and biological information content of HTS and profiling data constitutes one of the foundations of information-intensive molecular pharmacology and chemogenomics [76][77]. To cope with these expectations, NIBR collaborates with the bioinformatics company GeneData [78] in order to explore and implement a comprehensive software for data analysis integrating the following key aspects: 1) HTS and profiling data QC addressing process quality issues; 2) data normalization and standardization, including annotation; and 3) analysis of biological and chemical information content of profiling and primary HTS data. Based on field analysis and literature reports [79], HTS and profiling data analysis is within the pharmaceutical industry a neglected activity. Interestingly, some biotech companies like Cerep have built their business model not only based on intense screening and profiling, but also based on high-level data analysis [80]. Whereas the QC aspect is more tightly related to the screening and automation process itself [81], the data standardization and annotation, and especially the analysis of the chemical and biological information content are close to the data mining and compound selection steps where disease area chemists and biologists are involved. Appropriate assay annotation is invaluable for analyses across multiple assays in order, for instance, to detect screening technology or target family specific artifacts like frequent hitters [18][19]. Alternatively, such annotations enable previously known active sets to be subtracted from newly generated HTS hit lists, providing in this way a strategy to enhance the potential for discovery of truly novel chemotypes from HTS campaigns. Additional value of the annotations is recognized as being a basis for rational navigation systems for the fast growing number of data sets generated within the HTS and profiling factories and to allow category-dependent monitoring of hit rates. Within this perspective, a limited number of annotation categories were defined by the screeners and data analysis experts and include for example: 1) Organizational data (customer disease area, screening department, *etc.*); 2) target classes (*e.g.* molecular target, pathways, *etc.*) including for protein targets a gross classification of the target

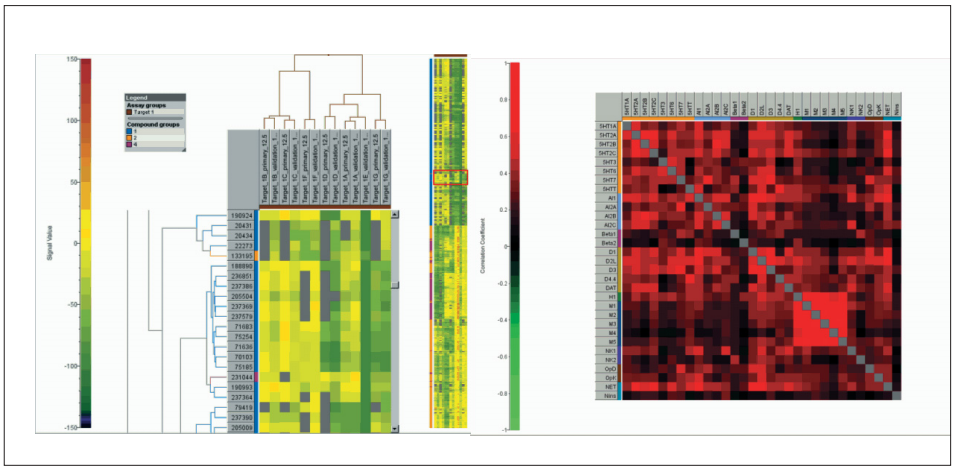

Fig. 4. Advanced data mining of large scale structure-activity bioprint-like and correlation matrices. The GeneData Sarileo software which NIBR co-developed allows the chemical and biological clustering and similarity analysis of large scale structure-activity matrices as shown in the left panel for primary and validation datasets of multiple HTS assays. The primary and validation data sets for each target are more similar than the cross-target similarities. The correlation analysis of validated IC50 profiling data of a set of NIBR reference compounds on a panel of GPCR targets is shown in the right panel. The molecular recognition between a number of these receptors is clearly correlated to what is important information for the design of selective compounds and for the design of safety pharmacology studies. The screenshots were made using the GeneData Sarileo software. For further details see [78].

families (*e.g.* GPCR, Ion channel, *etc.*); 3) assay class (*e.g.* physicochemical, biochemical, *etc.*) and assay types (*e.g.* agonist, antagonist, *etc.*); 4) readout type (*e.g.* fluorescence, radioactivity, *etc.*); and 5) key experimental parameters (*e.g.* concentrations, incubation times, *etc.*). The compilation and analysis of large-scale structure–activity matrices is the goal of profiling activities and is also desirable for primary HTS data. Analyses pioneered by the Weinstein group at the NCI [76] and by Cerep [80–83] can here be considered as model systems (Fig. 4).

Essential chemical and biological analyses allowing compound selections and advanced data mining include for instance: 1) Selection of a data subset for analysis accessible *via* a standardized data navigation system; 2) interactive visualization of structure–activity arrays together with chemical structures; 3) compound filtering based on property ranges and occurrence of specified substructures; 4) chemical clustering of arrays, with the possibility to define subgroups of compounds for further analysis. The functionality is especially important within the analysis of primary HTS hit lists for the selection of compounds for validation screening and the reporting of final HTS hit lists; 5) chemical similarity searching starting from individual compounds or from compound clusters; 6) biological activity or property-based clustering of arrays for the identification of subgroups of compounds with similar mode-of-action or property profiles, as well as for the identification of existing profile classes *per se*. This functionality is especially important within the analysis of primary HTS hit lists for the selection of compounds for validation screening using multiple read-outs or based on additional screening data; 7) biological activity or property-based similarity searching for the identification of compounds with a specified profile; and 8) correlation analysis between assays. The correlation between assays is directly related to methods for the extraction of both frequent hitters and privileged scaffolds, the former being uninteresting compounds which should be eliminated and the latter being of interest for the enrichment of the corporate collection.

One of the main open questions in HTS data analysis focuses on the value of primary HTS data for the generation of Quantitative Structure–Activity Relationship (QSAR) models including the prediction of selectivity profiles and activity patterns across multiple assays/targets [85][86]. Because pharmacological efficacy and potency are not necessarily correlated, it might indeed be possible that such analyses are only meaningful for accurate binding data obtained from dose–response curve experiments and not for functional data obtained from single dose screening [87]. The analysis of the relationship between chemical clusters and biological profiles is of particular interest in the field of chemogenomics [88]. Interesting scientific questions address the relevance of chemical classes [82][83]: To what extend do similar compounds or compounds of a given chemical class show a conserved biological profile? Or, *vice-versa*: What is the diversity of compounds with a specific biological profile?

Despite these emerging advanced data analyses, today the basic HTS data analysis process is in most cases single-project focused with the objective to assist the selection of compounds from the primary HTS hit list for dose dependent validation screening. Depending on the size of the primary HTS hit list, this triage is basically done according to two different scenarios. 1) Reducing the size if the primary hit list is, as in many functional signaling and reporter-gene assays, too large compared to the naturally limited resources for validation screening, and 2) increasing the size of the primary hit list if it does not reach this limit, as it is the case for many cell-free biochemical assays. In both scenarios, the objective is to select and to enrich from a given primary HTS data set those compounds which have the best potential to become hit-to-lead compounds and to explore at maximum the chemical diversity represented in a primary hit list and other sources of information relevant for this specific screening campaign. As compound selection and filtering is a subject of intense scientific debate, the computational analysis process uses in a first step data pipelining tools to annotate the different decision criteria to the compounds (Fig. 5).

The annotation criteria are of diverse nature. Because of the legacy of the screening collection, compounds violating the standard substructure filters used in the design of the newer screening sets need to be applied. In addition, project specific substructure, scaffold and physicochemical filters are applied to the primary hit list in order to maximize the chemical attractiveness of the resulting hit list. Based on the chemist-dependent information of chemical attractiveness, Bayesian classifiers or other machine learning techniques can be applied to translate this information into predictive computational models [82][83]. In a similar manner, empirical information about the promiscuity or cell toxicity of the hitters can be integrated using reference lists compiled over the years by the individual screening labs for assays of the same format or same target family. Input from maximum common substructure clustering methods is used to track quickly chemical scaffolds that are over representative in a hit list [89][90]. Another use of clustering is to reduce large hit lists by cherry picking a representative set from each cluster preserving the most active compounds. The summary of the different annotation criteria can then be used to qualify the chemical and biological hit attractiveness using simple additive point-based scoring schemes. The annotated and scored primary hit lists are then discussed within the project team for a final decision. Hit list enrichment can be performed using several techniques like the already mentioned homology-based similarity searching using known reference

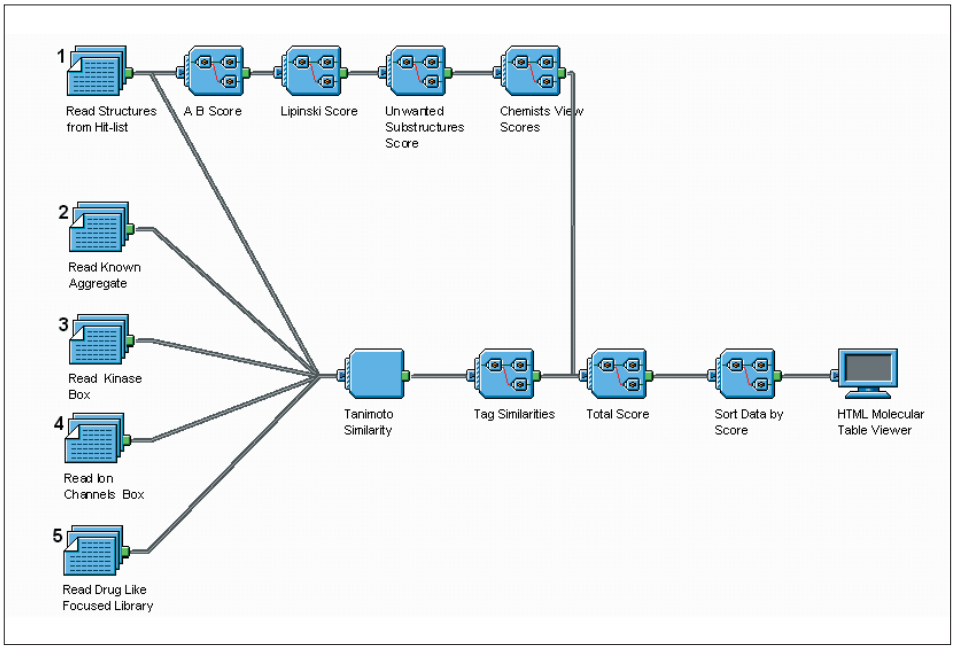

Fig. 5. Usage of Pipeline Pilot data processing software for primary hit list triage. In the shown computational protocol, the hit list is first annotated based on so-called A and B exclusion substructure filters. Lipinski-based properties and chemist specific substructure annotations are then added. In the next steps, the Tanimoto similarity compared to a number of reference compound lists is annotated. All annotations are then summarized into an overall score and the fully annotated primary hit list is then discussed by the project team in order to make the selection of compounds to be validated. For further detail on the Pipeline Pilot software see reference [41].

compounds and HTD or pharmacophore searching when a 3D model of the target binding site is available. More recently, predictive QSAR modeling techniques are used for the prediction of false positives and negatives from primary HTS data [85][86]. Follow-up additions using similarity and pharmacophore-based searching using the validated screening results are then performed in most cases within the disease area. The importance to trace the triage annotations and decisions as derived data is recognized as being of capital interest for the transparency of the HTS hit identification process. A major challenge for HTS data analysis is to keep up with the rapid evolution of HTS techniques and read-outs which, in many cases, require specific novel analysis approaches [91–93].

5. Conclusion

The selection of small molecular weight compounds for screening and hit-to-lead activities is of central importance for drug discovery. As presented in this article, the role of molecular informatics within the NIBR Discovery Technologies is to provide an *in silico* platform for the integrated generation and analysis of data and knowledge relevant for high-throughput science driven lead finding. To further improve the impact of this industrial approach, the experimentally and *in silico* generated data and knowledge will need to be further integrated together with advanced data mining technologies accessible to the drug discovery experts in the disease areas. Consequently, in order to maximize the added value for the overall discovery and optimization process, our current efforts focus on the consistent chemical and biological annotation of the data and decisions, aiming at the development of drug discovery ontologies at the genome level. Such advanced knowledge-based [2][94] systems are expected to complement the expertise of the drug discovery chemist and biologist and should especially allow us to learn more efficiently from the past experiences in order to explore more quickly the biology of novel targets and pathways, and to contribute by the development of mechanism and knowledge-based medicines to improve the overall productivity of the industrial pharmaceutical research.

Acknowledgments

Drs S. Heyse (GeneData, Basel), W. Andreoni and A. Curioni (IBM Research, Zürich), and Prof R. Stoop (ETH Zürich), P. Willett (University of Sheffield, UK), and R. Efremov (Russian Academy of Sciences, Moscow), and numerous colleagues in the NIBR IT, HTS and chemistry groups are acknowledged for various discussions. Parts of the work reported herein were done within the frame of the Swiss KTI research project 'Information-based Approaches in Drug Design', whose support is gratefully acknowledged.

Received: July 2, 2004

[1] R. Glen, *Chem. Comm.* **2002**, *23*, 2745.
[2] P. Gund, E. Maliski, F. Brown, *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 283.
[3] R.L. Strausberg, S. Schreiber, *Science* **2003**, *300*, 294.
[4] *www.inpharmatica.com*.
[5] D.A. Debe, K. Hanbly, *Curr. Drug Discov.* **2004**, *3*, 15.
[6] M. Bredel, E. Jacoby, *Nat. Rev. Genetics* **2004**, *5*, 262.
[7] 'Virtual Screening for Bioactive Molecules', Eds. H.-J. Böhm, G. Schneider, Wiley-VCH, Weinheim, **2000**.
[8] J. Bajorath, *Nat. Rev. Drug Discov.* **2002**, *1*, 882.
[9] A.L. Hopkins, C.R. Groom, *Nat. Rev. Drug Discov.* **2002**, *1*, 727.
[10] R.P. Sheridan, S.K. Kearsley, *Drug Discov. Today* **2002**, *4*, 903.
[11] K.V. Balakin, E.S. Tkachenko, S.A. Lang, I. Okun, A.A. Ivashchenko, N.P. Savchuk. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1332.
[12] P. Schneider, G. Schneider, *QSAR Comb. Sci.* **2003**, *22*, 713.
[13] M. von Korff, M. Steger, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1137.
[14] A. Schuffenhauer, P. Floersheim, P. Acklin, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391.
[15] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177.
[16] J. Hert, P. Willett, D. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *Org. Biomol. Chem.* **2004**, in press.
[17] G.M. Rishton, *Drug Discov. Today* **2002**, *8*, 86.
[18] O. Roche, P. Schneider, J. Zuegge, W. Guba, M. Kansy, A. Alanine, K. Bleicher, F. Danel, E.-M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, M. Dillon, E. Sjögren, N. Fotouhi, P. Gillespie, R. Goodnow, W. Harris, P. Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann, G. Schneider, *J. Med. Chem.* **2002**, *45*, 137.
[19] J. Seidler, S.L. McGovern, T.N. Doman, B.K. Shoichet, *J. Med. Chem.* **2003,** *46*, 4477.
[20] R. Crossley, *Curr. Top. Med. Chem.* **2004**, *4*, 581.
[21] P. Jimonet, R. Jäger, *Curr. Opin. Drug Discovery Dev.* **2004,** *7*, 325.
[22] A. Schuffenhauer, J. Zimmermann, R. Stoop, J.J. van der Vyver, S. Lecchini, E. Jacoby, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947.
[23] E. Jacoby, A. Schuffenhauer, P. Acklin, in 'Chemogenomics in Drug Discovery – A Medicinal Chemistry Perspective', Volume 22, 'Methods and Principles in Medicinal Chemistry', Eds. R. Mannhold, H. Kubinyi, G. Folkers, Wiley-VCH, Weinheim, **2004,** pp. 139–166.
[24] A. Schuffenhauer, E. Jacoby, *BioSilico* **2004**, *2*, 190.
[25] D.E. Root, S.P. Flaherty, B.P Kelley, B.R. Stockwell, *Chem. Biol.* **2003**, *10*, 881.

[26] R.A. Goodnow, W. Guba, W. Haap, *Comb. Chem. High Throughput Screen.* **2003**, *6*, 649.
[27] M.D. Burke, S.L. Schreiber, *Angew. Chem. Int. Ed.* **2004**, *43*, 46.
[28] A. Schuffenhauer, M. Popov, U. Schopfer, P. Acklin, J. Stanek, E. Jacoby, *Comb. Chem. High Throughput Screen.* **2004**, in press.
[29] P.S. Charifson, W.P. Walters, *J. Comput.-Aid. Mol. Design* **2002**, *16*, 311.
[30] R.D. Clark, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181.
[31] G.M. Downs, J.M. Barnard, *Rev. Comput. Chem.* **2002**, *18*, 1.
[32] *http://www.bci.uk.com.*
[33] T. Ott, A. Kern, A., Schuffenhauer, M. Popov, P. Acklin, E. Jacoby, R. Stoop, *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1358.
[34] H.O. Villar, R.T. Koehler, *Mol. Divers.* **2000**, *5*, 13.
[35] R. Nilakantan, D.S. Nunn, *Drug Discov. Today* **2003**, *8*, 668.
[36] G. Harper, S.D. Pickett, D.V. Green, *Comb. Chem. High Throughput Screen.* **2004**, *7*, 63.
[37] P. Watson, P. Willett, V.J. Gillet, M.L. Verdonk, *J. Comput. Aided Mol. Des.* **2001**, *5*, 835.
[38] W.H.B. Sauer, M.K. Schwarz, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987.
[39] G. Müller, *Drug Discov. Today* **2003**, *8*, 681.
[40] *www.chemnavigator.com.*
[41] Pipeline Pilot Software, SciTegic™, *www.scitegic.com.*
[42] R. Abagyan, M. Totrov, *Curr. Opin. Chem. Biol.* **2001**, *5*, 375.
[43] R.D. Taylor, P.J. Jewsburyet, J.W. Essex, *J. Comput. Aided Mol. Des.* **2002**, *16*, 151.
[44] G. Schneider, H.J. Böhm, *Drug Discov. Today.* **2002**, *7*, 64.
[45] N. Brooijmans, I.D. Kuntz, *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335.
[46] H.-J. Böhm, M. Stahl, *Rev. Comput. Chem.* **2002**, *18*, 41.
[47] M. Jacobsson, P. Liden, E. Stjernschantz, H. Bostrom, U. Norinder, *J. Med. Chem.* **2003**, *46*, 5781.
[48] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P.Francis, P.S. Shenkin, *J. Med. Chem.* **2004**, *47*, 1739.
[49] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J.L. Banks, *J. Med. Chem.* **2004**, *47*, 1750.
[50] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R.J. Taylor, *J. Mol. Biol.* **1997**, *267*, 727.
[51] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470.
[52] J. Fernandez-Recio, M. Totrov, R. Abagyan, *Proteins* **2003**, *52*, 113.
[53] *www.accelrys.com.*
[54] *www.tripos.com.*
[55] E. Vangrevelinghe, K. Zimmermann, J. Schoepfer, R. Portmann, D. Fabbro, P. Furet, *J. Med. Chem.* **2003**, *46*, 2656.
[56] J.L. Jenkins, R.Y. Kao, R. Shapiro, *Proteins* **2003**, *50*, 81.
[57] R.C. Stevens, S. Yokoyama, I.A. Wilson, *Science* **2001**, *294*, 89.
[58] J. Kopp, T. Schwede, *Pharmacogenomics* **2004**, *5*, 405.
[59] C. Bissantz, A. Logean, D. Rognan, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1162.
[60] M.L. Lamb, K.W. Burdick, S. Toba, M.M. Young, A.G: Skillman, X. Zou, J,R. Arnold, I.D. Kuntz, *Proteins* **2001**, *42*, 296.
[61] G.P. Vigers, J.P. Rizzi, *J. Med. Chem.* **2004**, *47*, 80.
[62] C. Bissantz, P. Bernard, M. Hibert, D. Rognan, *Proteins* **2003**, *50*, 5.
[63] X. Chen, C.Y. Ung, Y. Chen, *Nat. Prod. Rep.* **2003**, *20*, 432.
[64] N. Paul, E. Kellenberger, G. Bret, P. Mueller, D. Rognan, *Proteins* **2004**, *54*, 671.
[65] X. Fradera, J. Mestres, *Curr. Top. Med. Chem.* **2004**, *4*, 687.
[66] S. J. Teague, *Nat. Rev. Drug Discov.* **2003**, *2*, 527.
[67] H. Claussen, C. Buning, M. Rarey, T. Lengauer, *J. Mol. Biol.* **2001**, *308*, 377.
[68] J.A. Erickson, M. Jalai, D.H. Robertson, R.A. Lewis, M.J. Vieth, *J. Med. Chem.* **2004**, *47*, 45.
[69] J. Shen, X. Xu, F. Cheng, H. Liu, X. Luo, J. Shen, K. Chen, W. Zhao, X. Shen, H. Jiang, *Curr. Med. Chem.* **2003**, *10*, 2327.
[70] J. Aqvist, V.B. Luzhkov, B.O. Brandsdal, *Acc. Chem. Res.* **2002**, *35*, 358.
[71] R.C. Rizzo, M. Udier-Blagovic, D.P. Wang, E.K. Watkins, M.B. Kroeger-Smith, R.H. Smith, J. Tirado-Rives, W.L. Jorgensen, *J. Med. Chem.* **2002**, *45*, 2970.
[72] J. Wang, P. Morin, W. Wang, P.A. Kollman, *J. Am. Chem. Soc.* **2001**, *23*, 5221.
[73] T. Mordasini, A. Curioni, R. Brusi, W. Andreoni, *ChemBioChem* **2003**, *4*, 155.
[74] W.P. Walters, M. Namchuk, *Nat. Rev. Drug Discov.* **2003**, *2*, 259.
[75] K.H. Bleicher, H.-J. Böhm, K. Müller, A.I. Alanine, *Nat. Rev. Drug Discov.* **2003**, *2*, 369.
[76] J.N. Weinstein, T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace Jr., K.W. Kohn, T. Fojo, S.E. Bates, L.V. Rubinstein, N.L. Anderson, J.K. Buolamwini, W.W. van Osdol, A.P. Monks, D.A. Scudiero, E.A. Sausville, D.W. Zaharevitz, B. Bunow, V.N. Viswanadhan, G.S. Johnson, R.E. Wittes, K.D. Paull, *Science* **1997**, *275*, 343.
[77] M.F. Engels, *Ernst Schering Res Found Workshop,* **2003**, 87.
[78] *www.genedata.com.*
[79] P. Gedeck, P. Willett, *Curr. Opin. Chem. Biol.* **2001**, *5*, 389.
[80] *www.cerep.fr.*
[81] C. Brideau, B. Gunter, B. Pikounis, A. Liaw, *J. Biomol. Screen.* **2003**, *8*, 634.
[82] D. Horvath, C. Jeandenans, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680.
[83] D. Horvath, C. Jeandenans, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691.
[84] C.M. Krejsa, D. Horvath, S.L. Rogalski, J.E Penzotti, B. Mao, F. Barbosa, J.C. Migeon, *Curr. Opin. Drug Discov. Devel.* **2003**, *6*, 470.
[85] A. E. Klon, M. Glick, M. Thoma, P. Acklin, J.W. Davies, *J. Med. Chem.* **2004**, *47*, 2743.
[86] T. Lengauer, C. Lemmen, M. Rarey, M. Zimmermann, *Drug Discov. Today* **2004**, *9*, 27.
[87] D. Colquhoun, *Br. J. Pharmacol.* **1998**, *125*, 924.
[88] M. Vieth, R.E. Higgs, D.H. Robertson, M. Shapiro, F.A. Gragg, H. Hemmerle, *Biochim. Biophys. Acta* **2004**, *11*, 243.
[89] C.A. Nicolaou, S.Y. Tamura, B.P. Kelley, S.I. Bassett, R.F. Nutt, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069.
[90] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower Jr., *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302.
[91] M. Entzeroth, *Curr. Opin. Pharmacol.* **2003**, *3*, 522.
[92] *http://www.dtp.nci.nih.gov*
[93] *http://www.speroid.ncifcrf.gov*
[94] J. Mestres, *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 304.