Drugs of the Future 1999, 24(2): 177-190
Copyright © 1999 PROUS SCIENCE
CCC: 0377-8282/99

Addressing the problem of molecular diversity

Uwe Eichler, Peter Ertl, Alberto Gobbi and Dieter Poppinger*

Research Computing, R-1007.8.04, Novartis Crop Protection AG, CH-4002 Basel, Switzerland.
*Correspondence

CONTENTS

Summary

Because the capacity of compound aquisition and screening in agrochemical research continues to be limited, intelligent compound selection strategies must be used. These “diversity methods” can be used to ensure that only sufficiently novel compounds are screened, to focus attention on compounds which have a high probability of being active and to refine initial leads by a rational process. These methods are briefly reviewed and illustrated by tools and examples from Novartis Crop Protection research.

Introduction

To speed up the discovery of new active substances, most agrochemical and pharmaceutical companies have set up high throughput screening (HTS) programs. Compound input is from natural products, chemistry projects aimed at filling out “holes” in competitor patents, lead optimization projects, isolated side products from laboratory or scale-up syntheses, screening samples which are acquired from academic and commercial sources and high speed synthesis (HSS). The *in vivo* HTS program at Novartis Crop Protection differs fundamentally from the pharmaceutical industry. Firstly, compounds are tested on whole organism assays which are very close to the actual targets (plants, fungi, insects). This limits the annual throughput to the order of 100,000, compared to the several million single compounds and mixture samples which are possible in pharmaceutical HTS programs. Secondly, and also in contrast to pharmaceutical screening, a constant stream of new compounds is tested on a fixed set of standard assays. Our current HSS/HTS discovery pipeline has two major bottlenecks: budgetary restrictions on the number of compounds which can be purchased and the compound handling and screening capacity. For example, there are probably at least 2 million compounds which can be bought from commercial sources at a price of US $10-150 per sample. It is not possible to purchase and screen them all. The potentially enormous output of combinatorial synthesis (1) poses even bigger challenges. Even a rather simple combinatorial library with three variable positions can, in principle, yield 27 million compounds, if one conservatively assumes that 300 different reagents are available for each position. For both sources of compounds for screening, it is therefore necessary to select those subsets which promise to contain the largest number of interesting leads.

There are many issues that have to be considered here, such as covering the synthetically accessible “chemical space” as completely as possible, avoiding areas already covered by competitor patents, forcing lead compounds to be significantly different from each other and from what has already been discovered, incorporating pharmacophoric structural motifs in order to increase the hit rate, or optimizing physical properties which will affect uptake or degradation in the field. We find it convenient to discuss these under the general heading of “diversity”, although not all are related to the question of structural diversity as such, but rather to the general problem of how to design optimal subsets of compounds for lead discovery and optimization, and whether this is worthwhile doing at all.

In the following, we will briefly review the field, with a particular emphasis on methods related to current practice at Novartis Crop Protection. For a more exhaustive review, a number of recent publications (2-4) can be consulted. In particular, we will discuss methods to increase the novelty of potential lead compounds and methods to increase the chances of finding active compounds (the hit rate). We will not cover the algorithms or their implementation in any detail; this will be the subject of a subsequent article. Rather, we will provide an illustration of the various techniques and tools which are used in a large agrochemical company to address the problem of molecular diversity.

Methods to increase the novelty of potential lead compounds

The novelty of lead compounds has several facets. In a chemical sense, compounds are novel if they are structurally sufficiently different from what has been screened before or will be screened at the same time, or what has been described in the relevant patent literature. We call a set of such molecules "structurally diverse". More precisely, sets of structures can be "intrinsically diverse" or "diverse with respect to a reference set". For example, it may appear attractive to purchase a set of 10,000 screening substances which the supplier advertises as being maximally different from each other (intrinsically diverse), according to one of many possible diversity criteria. However, if most of these compounds happen to be from a chemical class which has been investigated in detail in earlier in-house screening programs, or is covered by a competitor's patent, then there are probably more promising choices one could make. By the same token, when planning the synthesis of a combinatorial library, one should not only evaluate the intrinsic diversity of the many possible libraries, but also their diversity relative to existing compound collections.

At least as important as these structural considerations is biological novelty, *i.e.*, whether a new compound also exhibits a new mode of action. Unfortunately, it is not possible to predict this with any degree of reliability. Therefore, we concentrate on the structural aspects of the problem.

Structural novelty

In principle, structural novelty is easy to evaluate since it involves nothing more than comparing all potential lead compounds with each other and with compounds in proprietary and public databases. In practice, there are some problems associated with the size of the necessary computations and also with the exact meaning of "sufficiently different". Thus, much of the recent technical work in the diversity field has been concerned with inventing efficient methods to store and compare molecules and with measuring and assessing the difference between molecules.

Table I: Structural overlap of various compound sources.

Source	Size	ALD	SPES	BRID	NCP
ALD	65,234				
SPES	153,741	3925			
BRID	216,331	8342	23,221		
NCP	498,141	9061	7004	7225	
TP	1,001,921	65,234	153,741	216,331	40,356

TP = total of unique commercial compounds on the Novartis Crop Protection files. NCP = set of compounds which have been screened in one of our HTS or standard screens.

One of the most efficient ways to store structural information for organic molecules is the SMILES line notation (5), where, *e.g.*, CCCO denotes *n*-propanol and CC(C)O *i*-propanol. At Novartis Crop Protection, we use this representation in our database of corporate and external structures. This database holds approximately 1.3 million proprietary Novartis compounds and 1 million unique substances (out of 1.8 million nonunique offerings) from various suppliers of screening samples. Table I shows the overlap of compound sets from different sources (6-8).

Extending novelty checks to patents is difficult since many generic structures claimed in the patent literature are infinite ("... where R is any organic residue...") and cannot be enumerated. More involved algorithms have been developed to deal with patent-related problems (9) but do not seem to be used widely.

Molecular similarity and diversity

One generally expects closely related compounds to have similar biological effects. If two compounds are "sufficiently similar", one would assay only one and infer the activities of the second from the measurements on the first. If the compound set to be screened contains many sufficiently similar molecules, or many compounds which are sufficiently similar to compounds in a previously screened reference collection, then it would suffice to screen a properly chosen subset. This subset should contain only compounds which are sufficiently dissimilar from any other compound in the screening set or the reference set, and should still contain all essential structure-activity information of the full set. To make this approach successful, one has to have a computable and biologically meaningful measure of molecular similarity and an algorithm to select the representative subset.

A large number of molecular similarity measures has been suggested in the literature (10); examples include measures based on chemical substructures, pharmacophore properties (11), interaction fields (12) and physicochemical properties (13). Recent work on validating the biological relevance of molecular similarity measures by determining how well these criteria differentiate between known active and inactive compounds (14, 15), has shown that simple substructure-based measures are to be preferred, since they are easier to compute and outperform other seemingly more sophisticated criteria.

Substructure-based similarity measures are usually based on structure fingerprints. A structure fingerprint is a compact record of structure-related molecular features which can be used to characterize a molecule, much like the fragmentation pattern in mass spectrometry is used to characterize a compound. Structure fingerprints are derived by systematically checking which substructures occur in the molecule, and stored in a computer as a string of 0 and 1 bits, where a 1 at the n-th position means that the n-th substructure is present. One popular type of 2D structure fingerprint is due to the Daylight group (16) and records the presence of all possible linear substructures up to a length of 7. Other fingerprint definitions are based on the structural fragments which are used in the MDL (17) or Tripos (18) chemical database systems.

The similarity of two molecules can be defined by counting the common occurrence of features and normalizing this count such that it is between 0 (no common features) and 1 (identical molecules). One measure of this kind is the Tanimoto similarity index $T_{AB} = T_{BA} = n_{AB}/(n_A + n_B - n_{AB})$, where n_{AB} is the number of features present in both molecules A and B, and n_A and n_B are the numbers of features present in molecule A and B (19). A related measure is the substructure or subset similarity $S_{AB} = n_{AB}/n_A$, which is equal to 1 if A is a substructure of B (or all features of A occur in B), and < 1 otherwise.

The fastest way to assess the intrinsic diversity of a set of compounds is to count the number of different features occurring in any structure in the set. For example, Davies and Briant (20) have used the number of different 3-point pharmacophores to measure diversity. A histogram displaying molecular similarities of all nearest neighbor molecule pairs is, in principle, a better approach. Such similarity profiles have been widely used to visualize the diversity of screening compound collections.

Selection algorithms

Having established a method to measure structural similarity and the diversity of a set of compounds, the task remains of how to select a diverse subset of compounds. The most commonly used methods are dissimilarity selection, clustering and partitioning.

In the *dissimilarity selection method*, compounds are selected in such a way that the maximum similarity of a pair of molecules in the selected set is minimized (21). In our implementation, the method works as follows: (a) the first compound of N_{CAND} candidates is selected randomly, (b) the next compounds to be selected are those with the largest nearest neighbor distance to any of the already selected molecules, and (c) these steps are continued until N_{SEL} compounds are selected. This algorithm requires N_{CAND}*N_{SEL} distance evaluations and is feasible for candidate data sets of about 1 million compounds. The same algorithm allows the selection of a subset which is not only intrinsically diverse, but also diverse to a given reference set.

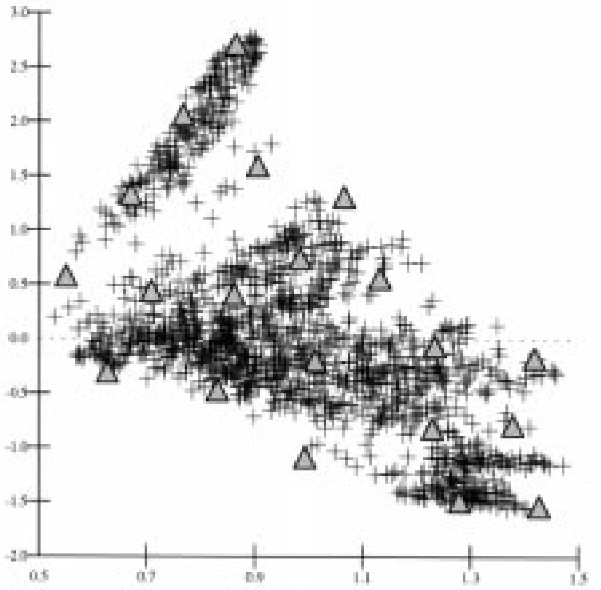
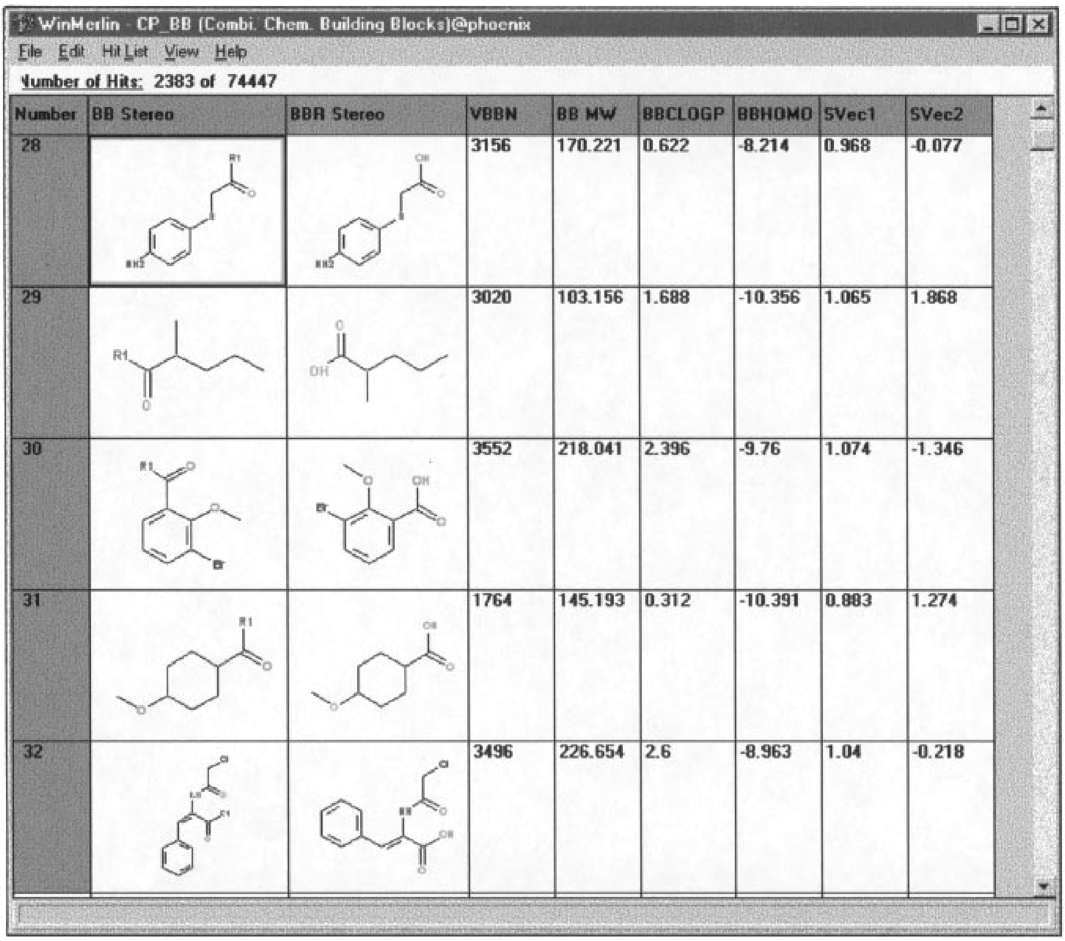


Fig. 1. Choosing 20 out of 1474 possible oxime building blocks for library synthesis by dissimilarity selection. The axes are the first two dimensions of a multidimensional scaling of the Tanimoto similarity matrix. Clusters represent broad structural classes (aliphatic, aromatic, heteroaromatic). The selection (triangles) overrepresents compounds at the periphery of the structure clusters, *i.e.*, compounds with unusual functionality.

Figure 1 shows that the dissimilarity method has a propensity to select outliers, unless special measures are taken. To suppress this tendency to select outliers we use a “selection distance” and accept compounds only if they are within this distance to any of the already selected molecules.

With *clustering methods* (22), one tries to sort similar molecules into groups. A smaller subset is constructed by selecting a representative molecule from each group. This subset is not necessarily maximally diverse, but it may be more representative for the entire collection than the results of dissimilarity selection. There are different clustering algorithms with different computational complexity and performance characteristics which will not be discussed here.

In the *partitioning method*, compounds are represented by points in a multidimensional space, the axes of which are defined by the parameters one chooses for measuring molecular similarity. This space is divided into cells of a fixed size, which implies that a decision has to be made on the significant resolution of parameters. A smaller subset of compounds can now be constructed by picking one representative compound from each cell. To construct a subset which does not overlap with a reference set (*e.g.*, the corporate compound collection), one would pick compounds only from those cells which are not occupied by reference compounds. The partitioning



Number of Hits: 2383 of 74447

Number	BB Stereo	BBR Stereo	VBBN	BB MW	BBCLOGP	BBHOMO	SVec1	SVec2
28			3156	170.221	0.622	-8.214	0.968	-0.077
29			3020	103.156	1.688	-10.356	1.065	1.868
30			3552	218.041	2.396	-9.76	1.074	-1.346
31			1764	145.193	0.312	-10.391	0.883	1.274
32			3496	226.654	2.6	-8.963	1.04	-0.218

Fig. 2. View onto the CICLOPS database of building blocks. Compounds can be selected manually or algorithmically, according to various physicochemical parameters or by distance in structure space.

method becomes impractical if the search space has too many dimensions, *i.e.*, if there are more than 6 or 7 different relevant parameters. To reduce the number of dimensions, principle component analysis and multidimensional scaling can be used (23).

To apply any of these methods to select subsets for screening, the required molecular properties (structure fingerprints, topological descriptions, pharmacophores, physicochemical data) have to be computed for all compounds which are of potential interest or have already been screened. This is feasible for sets of about 1 million compounds, as we shall discuss below.

To apply any of these methods to the design of diverse combinatorial libraries, molecular properties have to be computed for all compounds which are possible, given the synthesis protocol and the sets of available starting materials (the “virtual library”). Such virtual libraries can easily contain more than 10^{12} structures and are beyond direct computational evaluation. As a shortcut, one can design optimally diverse sets of reagents or building blocks (24) and assume that this will lead to an intrinsically diverse library. It is clear that this is an approximation (25). Whether this approximation has a measurable influence on the quality and number of hits which a combinatorial library produces is not known.

We have previously described the construction of CICLOPS, an integrated system to support the design and registration of combinatorial libraries at Novartis (26). In this system, we have implemented library design at the building block level. Specifically, CICLOPS automatically calculates a large number of structure-related and physicochemical parameters for any building block that is newly registered by a chemist. Principle component analysis and multidimensional scaling of these parameters is then used to define an orthogonal design space of low dimensions. The dissimilarity selection algorithm and others methods are used to select a subset of building blocks, which are structurally diverse or span the required range of physicochemical parameters as evenly as possible. This process is illustrated in Figure 2.

CICLOPS has been in routine use for several years and has been central to the combinatorial chemistry efforts at Ciba-Geigy and now Novartis. However, this is not because of the important role of library design methods, but because our chemists appreciate CICLOPS as a convenient tool for keeping track of the reagent stockroom, browsing through catalogs of available reagents, selecting building blocks by chemical criteria, and for enumerating combinatorial libraries into lists of molecules which are finally registered into the screening database systems. In comparison, library design methods are seldom used. We therefore have no empirical evidence that rational library design methods lead to more interesting combinatorial libraries than the manual approach currently used by our chemists.

Diversity and hit rate

It is not clear *a priori* whether the hit rate of a screening collection can be improved by diversity design. If a set of compounds is truly random, then a random subset has the same chance of containing active compounds as a set of the same size which was designed to be structurally diverse. The diverse set, by construction, will contain more different compounds, *i.e.*, be intrinsically more novel but will not contain more actives.

The screening collections which have been accumulated in large corporations are not random, but contain clusters of structurally related active compounds which reflect the "homing in" process of lead optimization projects. A random subset of these clustered collections will contain, on the average, more compounds from the well-examined active areas than a diverse subset. Screening a diverse subset against a new target may or may not produce more hits than a random subset, depending on whether or not the new target has significantly different structural requirements. It has been reported that rescreening diverse subsets of historical compound collections against new targets produces more (27, 28), or at least more interesting (29), hits than rescreening random subsets. However, there is no reason to expect the same effect for commercial compound collections, which are structured differently. We are, therefore, developing generally applicable, rule-based methods to bias subsets of compound collections towards active molecules.

Selection methods to increase the probability of finding active molecules (enrichment methods)

Applying rule-based methods to lead finding, either through compound acquisition or through high speed synthesis, has a fundamental limitation. Most rule-based models are local, *i.e.*, they cannot be transferred from one structure class to another and they are optimized for one particular biological test system. What is needed instead for lead finding are global structure-activity models which cover the entire space of synthetically reasonable compounds, can deal with several targets at a time and are based on parameters which can be computed without manual intervention for millions of candidate structures. The historical screening collections of major pharmaceutical and agrochemical companies are a rich source of data to develop such global models.

Of the multitude of molecular descriptors which have been used in structure-activity studies, physicochemical bulk parameters such as the octanol-water partition coefficient, molar refractivity and water solubility appear to be the most promising for our *in vivo* lead finding efforts, since they are related to transport properties. They can be computed rather easily using fragment contribution schemes (30, 31). Other parameters which we have used successfully in deriving local QSAR relations for agrochemicals are frontier orbital energies (32) to measure donor-acceptor properties, selected atomic charges (32) and aligned molecular fields (33). Other groups have attempted to relate the occurrence of molecular fragments to activity (34) and toxicity (35, 36), and compared the success rate of such schemes with that of human studies (37).

To derive and validate an activity model, a training set and a validation set of structures with activity data are needed. These are conveniently constructed by randomly dividing the entire historical screening collection (or a subset which has been made structurally representative by clustering or similar techniques) into two halves, a training set and a test set. The model is optimized to reproduce the selected activity data in the training set and preferably cross-validated by the leave-one-out or other methods (38). Testing the model against the validation set gives some indication of how well it will perform in identifying active compounds in different commercial collections or in virtual combinatorial libraries.

In principle, one could aim at deriving a separate model for each target of interest, a single model for all possible biological targets, or a small number of models which represent a group of related targets like all weeds. The latter may be a reasonable balance between the desire to construct simple and general models and the necessity to capture factors which differentiate between, *e.g.*, phytotoxicity and antifungal activity.

Iterative activity optimization

The QSAR-type approach described above uses a representative training set to derive a model which is then fixed and employed to select compounds from different sources. Alternatively, one can use an iterative model refinement approach. Given a set of weak lead structures, an initial model is constructed and used to propose the next set of structures which are synthesized and tested. The information from the biological tests is used to refine the model, which is then again used to propose the next round of synthesis and testing (Fig. 3).

Weber *et al.* (39) and Singh *et. al.* (40) have pioneered this approach, using a model consisting simply of a list of reagents which lead more often to active compounds than to inactives. Myers *et al.* (41) have outlined a more complex method based on 3D pharmacophores as models for receptor binding. We have developed a genetic optimization method based on 2D structure similarities (42) which is described below.

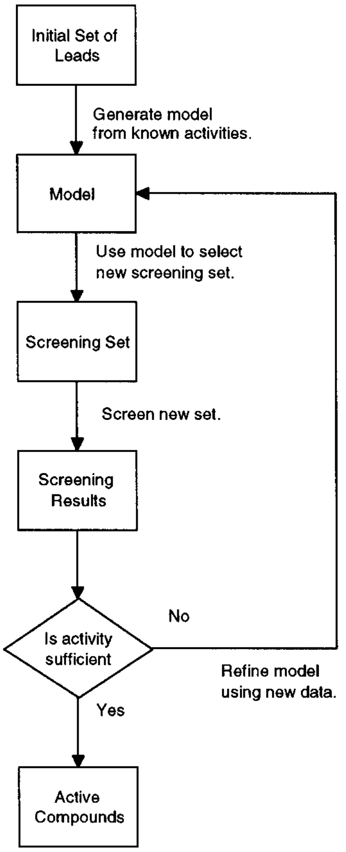


Fig. 3. The iterative model refinement approach.

Genetic algorithms are optimization methods which are known to be well suited to difficult optimization problems (43). The basic genetic algorithm is very simple. An initial population of solutions is chosen, either at random or based on existing information. The solutions are encoded into a “chromosome”, which is usually represented as a string of characters which describe the problem space. Solutions are changed by “mutation” (by applying random modifications to the chromosome string) or by “crossover” (combining two solution chromosomes according to some systematic recipe). A “fitness” score is then computed for the new solutions, and the fittest descendants are allowed to replace the weakest parents. In our approach, we use folded substructure fingerprints (16) as structure-related chromosomes and the experimental activity data as a fitness function. This method can be applied to select database subsets for screening and to optimize combinatodal libraries for activity against selected targets (44).

Applications

Selection of compounds for high throughput screening

In Novartis Crop Protection, we have built up a database containing more than 1 million unique compounds from external suppliers of screening compounds. Of these compounds, about 900,000 are different from the already known in-house compounds. In the project described here, 10,000 compounds were selected for the high throughput screening (HTS) program. The selected compounds should be equally distributed in chemical space, and should also be diverse with respect to compounds which had already been screened in Novartis Crop Protection, so that screening of similar compounds is avoided. Since it is known that the dissimilarity selection algorithm (19) initially selects outliers with undesirable structural characteristics, these were filtered out. We decided to remove all compounds (1579) with a molecular weight < 100 and > 1000 D, and all compounds (5679) with unusual (other than H, C, N, O, S, P, F, Cl, Br, I, Li, B, Al, Na, K, Ca, Mg) elements. The remaining 901,078 compounds defined the candidate set for diversity selection.

As a first step of the selection procedure, the nearest neighbor distances between in-house substances and compounds in the candidate set had to be determined. Since this is the most time-consuming part, requiring about 540 billion pairwise comparisons, this step was run in parallel on several Silicon Graphics Computers. It needed about 450 CPU hours to complete. A histogram for the nearest neighbor pair distances is shown in Figure 4. Most distances are in the range between 0.1 and 0.3, which means that most compounds offered by external suppliers are quite similar to our in-house structures and should probably not be considered for lead finding purposes.

For the diversity selection, different selection distances (see above) were applied. A selection distance of 1.0 (everything accepted) resulted in too many compounds with undesirable features, such as boron heterocycles, or compounds containing no carbon atoms (Fig. 5). A selection distance of 0.4 successfully removed such unreasonable compounds (Fig. 6).

Comparison of various enrichment methods

As discussed in the introduction, the goal of enrichment methods is to increase probability of selecting active

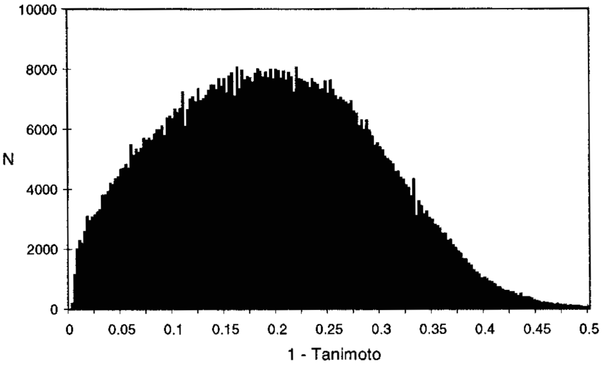


Fig. 4. Histogram displaying the nearest neighbor distances between third-party supplier substances and substances in the Novartis Crop Protection compound collection.

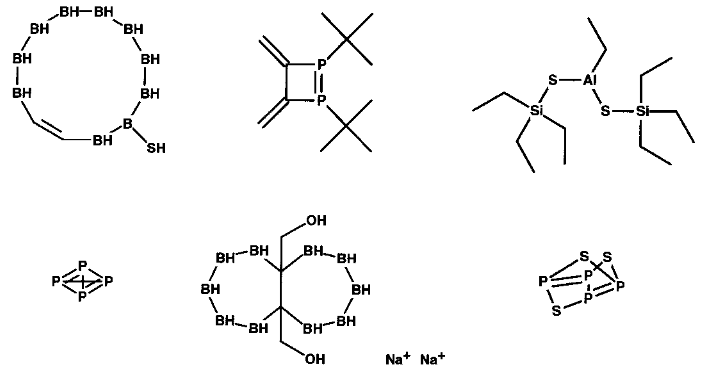


Fig. 5. Structures obtained by diversity-based selection with a selection barrier of 1.0. This procedure provides the most diverse compounds but at the same time emphasizes outliers.

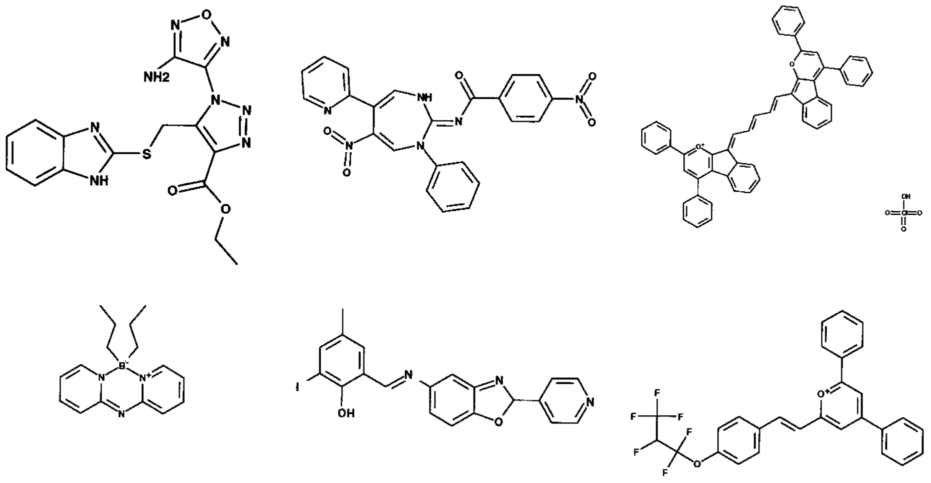


Fig. 6. Structures obtained by diversity-based selection with a selection barrier of 0.4. Outliers are avoided but the selected compounds are still diverse.

molecules from the available pool. In this section, we compare various enrichment strategies based on physicochemical parameters, simple constitutional descriptors and various structural fragments. A data set consisting of 51,093 molecules was created by selecting representative structures from the Novartis Crop Protection database. Out of this set, 8936 molecules (17.8%) are active. A molecule was considered to be active if it showed activity in any of the Novartis insecticidal screens. This is a challenge for classification schemes, since this definition of activity is very broad and covers many modes of action and various structural classes.

To classify molecules as active or inactive, an activity score is used. This is simply the sum of feature contributions which depend on how often these features (*i.e.*, particular values of physicochemical properties, number of halogens, particular fragments) occur in active or inactive compounds. If the presence of a particular feature does not relate to biological activity, then the percentage of molecules containing this feature should be the same for active and inactive subsets. Hence, the contribution of this feature to the activity score is zero. In some cases, however, the presence or absence of a feature is related to activity, and considerable deviations from this equal distribution occur. In these cases, the features have high positive (or negative) contribution to the activity score. To determine which sets of features are best, a measure suggested by Kearsley *et al.* (13) was used. This measure is based on a simulated screening experiment, where N compounds in a historical screening database are "tested" in order of their decreasing calculated activity score. The "test" is simply looking up the measured activity in the database. The initial enhancement is defined as the number of active molecules encountered in the first small subset of the database, divided by the number of actives expected by pure chance. For instance, if a 10,000 compound database contains 2000 active compounds, one would expect to find 200 actives after screening 1000 randomly chosen compounds. If, by applying a selection algorithm, one finds 800 actives instead, then the enrichment factor is 800/200 = 4 (denoted as A@1000 = 4). In this work we focus on the

Table II: Classification based on physicochemical properties.

Property	A@1000
LogP	1.69
Molar refractivity	1.27
Water solubility	1.62
Vapor pressure	1.26
Dipole moment	1.23
HOMO energy	1.39
LUMO energy	1.60
Molecular weight	1.25
LogP + LUMO	2.21
All 8 properties combined	2.33

Table III: Classification based on simple constitutional descriptors.

Key	A@1000
# Halogens	1.98
# Ring bonds	1.82
# Heteroatoms (not halogen)	1.80
# HB acceptors	1.46
Best 4 keys combined	2.22

comparison of various classification schemes and not on the quantification of performance. Therefore, no cross-validation was used (*i.e.*, the classification was tested on the same data set which was used for the generation classification rules).

At first, a classification based on physicochemical properties was tested. The range of a particular property was divided into 40 bins and the number of active and inactive molecules with property values in particular bins was compared. The resulting classifications were unsatisfactory. Of eight properties tested, only logP, water solubility and LUMO energy classified significantly better than expected by chance (Table II).

Classifications based on simple constitutional descriptors (number of atoms or bonds of a particular type, number of rings, aromatic atoms, H-bond donors and acceptors, *etc.*) led to a slightly better separation than classifications based on physicochemical parameters. Data for the best four keys are shown in Table III.

It is encouraging that the keys identified as the most important have clear physical meaning related to hydrophobicity (# halogens), structural flexibility (# ring bonds) or ligand-receptor interactions (# HB acceptors, # heteroatoms).

The best results were obtained by using simple fragments as classification criteria (Table IV). Two types of more generic fragments were considered: linear fragments and atoms with environment (atomic hose codes (45)). In Table IV, the initial enrichments for particular fragment sizes and also the number of different fragments contributing to the overall activity scores are shown.

From Table IV one can see that the branched fragments (hose codes) are performing slightly better than linear fragments. Note also that the maximum possible value of A@1000 for this data set is 5.75.

Table IV: Classification based on linear fragments and hose codes.

Fragments	A@1000	# scores
Linear, length 1	1.88	13
Linear, length 2	2.53	55
Linear, length 3	3.54	227
Linear, length 4	3.50	720
Linear, length 5	4.08	1738
Linear, length 6	4.31	3072
Linear, length 7	4.41	4083
Linear, length 8	4.41	4308
Hose, level 1	1.88	13
Hose, level 2	4.02	470
Hose, level 3	4.71	2229
Hose, level 4	4.76	2528

Our results are consistent with previous findings of other authors (46) who found that MACCS keys (simple two or three atomic fragments used by MDL software) (17) classify somewhat better than fingerprints based on linear fragments, and better than physicochemical parameters or 3D structural data. A more detailed comparison of classification schemes, including cross-validation, will be described elsewhere.

Genetic optimization for activity

A second approach which is currently used at Novartis Crop Protection is activity optimization using a genetic algorithm (GA), as described in the introduction. Our GA combines directed optimization and diversity selection. The optimization is directed because it combines the features of two active compounds (the parents) in the crossover step. The diversity component is introduced by the mutation operator, which pushes the search out of local minima and thereby increases the probability to find global minima.

Our GA uses a small population of parents to build a local model and refines this over a number of iterations. Compared with the classification methods described in the previous section, this has some advantages and disadvantages. Because the model is refined locally, it should be more predictive in this local region than a model which tries to cover a larger variety of structure types. In addition, the model is adjusted (and hopefully improved) throughout the entire optimization process, so there is a chance that more than one class of active compounds is found. On the other hand, this iterative process also has a disadvantage: each iteration cycle consisting of selection, acquisition and screening of the compounds must be finished before the next cycle can be started. This causes logistical and timing problems. Since at least about 20 cycles seem to be necessary to find highly active compounds, the method is feasible only if each cycle can be completed in less than 3 weeks.

This iterative screening process has been simulated using various historical data sets and produced very good results. As an illustration, we use a data set from the

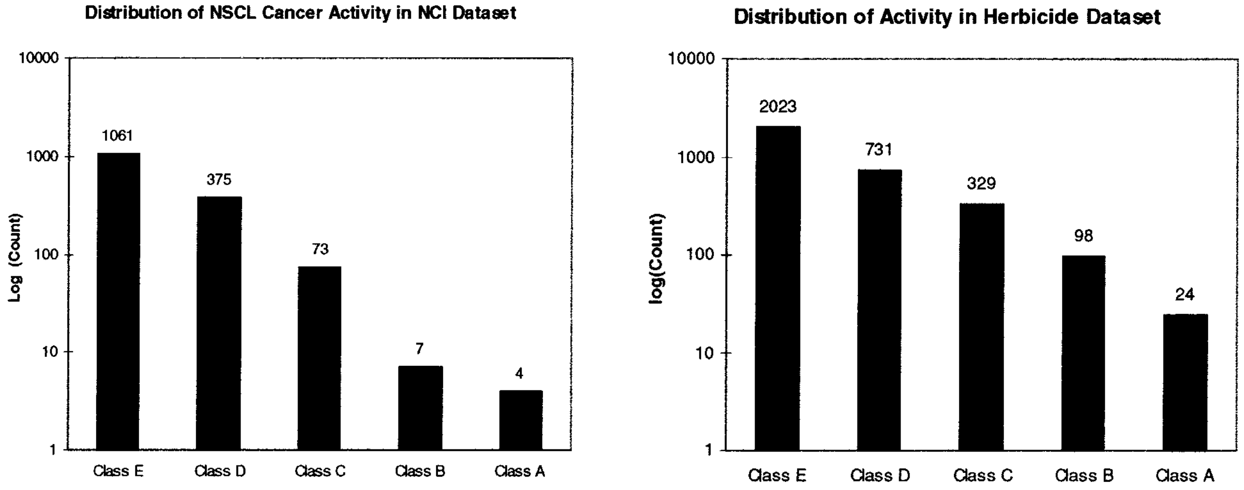


Fig. 7. Distribution of compounds in the five activity classes for the NCI and herbicidal data sets.

National Cancer Institute (47), which contains "Non-Small Cell Lung Cancer" assay data and an internal data set with activities from our herbicidal screening.

The NCI set consisted of 19,596 compounds with GI_{50} (equivalent to IC_{50}) values. Of these, 18,076 had $\log(GI_{50})$ values smaller than 5.55 and were classified as inactive. The remaining compounds were assigned to five classes of activity by dividing the $\log(GI_{50})$ range 5.55-13 into equally spaced bins. The herbicide data set consisted of 76,889 diverse compounds from our in-house screening database. Of these, 3205 compounds were classified as active and also divided into equally spaced activity bins. The activity distribution is illustrated in Figure 7.

For both data sets, 20 inactive compounds were selected at random to define the first set of parents. In each iteration, the most active 20 compounds screened thus far were used as a new parent set. From this set, 20 parent fingerprints were constructed by crossover between 20 randomly selected pairs of parents. Mutation was introduced by adding 2 random compounds to the parent set, before selecting the pairs for crossover. A new set of screening candidates was then constructed by retrieving those 20 molecules from the data set which have the highest substructure similarity to the parent fingerprints. The 20 candidates were then removed from the candidate data set and added to the set of screened compounds according to their activity.

Typical traces of this process can be seen in Figure 8. Here, the percentage of compounds successfully identified in each activity class is plotted against the

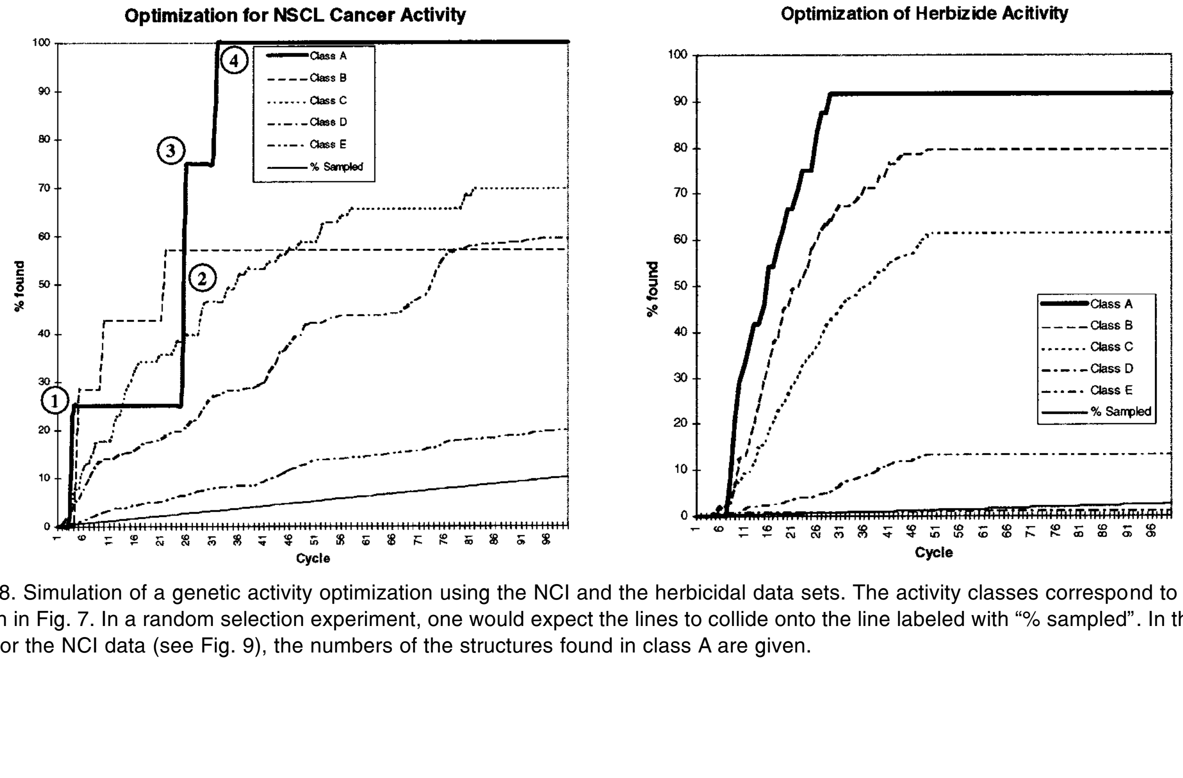


Fig. 8. Simulation of a genetic activity optimization using the NCI and the herbicidal data sets. The activity classes correspond to those given in Fig. 7. In a random selection experiment, one would expect the lines to collide onto the line labeled with "% sampled". In the figure for the NCI data (see Fig. 9), the numbers of the structures found in class A are given.

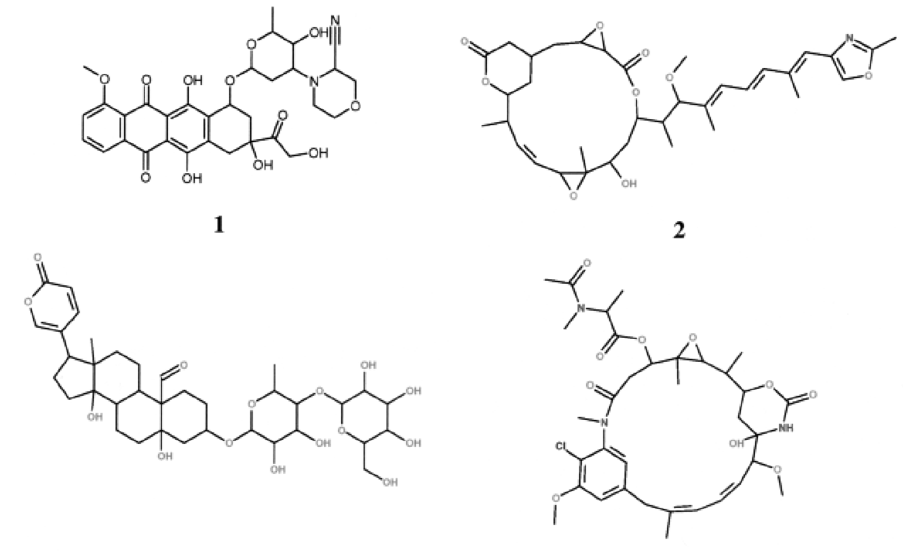

1 2 3 4

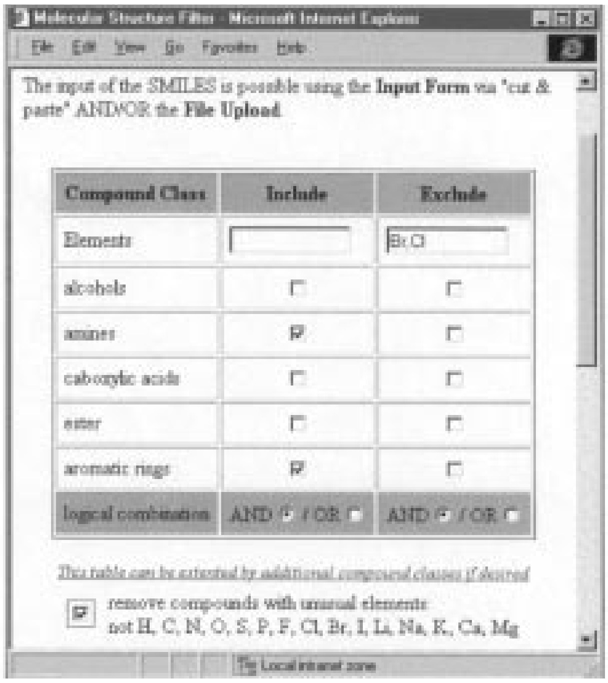
Fig. 9. Structures from class A found in the simulation using the NCI data set.

optimization cycle. As can be seen, the number of identified compounds in the most active class (A) begins to increase just after a few cycles. After about 30 cycles, most of the compounds from this class have been found, as well as more than 50% of the compounds from the less active class B. This might be compared to what one would expect from a random selection experiment (Fig. 8, the solid line labeled "% sampled").

Figure 9 shows the NCI structures from the class with the most active compounds. It is evident that the GA is indeed able to find compounds which belong to different structural classes, and we are currently using this approach to select compounds for one of our insecticidal screens from a number of preferred suppliers.

Tools in Novartis Crop Protection

Because there will always be selection criteria which are too difficult to formalize for general use, it is important that chemists are involved directly in selecting compounds for screening and in designing combinatorial libraries. Therefore, we have written a number of simple, web-based tools for compound selection which are available to all chemists at Novartis Crop Protection via our intranet.


Molecular Structure Filter - Microsoft Internet Explorer
File Edit View Go Favorites Help
Local intranet zone

The input of the SMILES is possible using the **Input Form** via "cut & paste" AND/OR the **File Upload**.

Compound Class	Include	Exclude
Elements		Br,Cl
alcohols	□	□
amines	☑	□
carboxylic acids	□	□
ester	□	□
aromatic rings	☑	□
logical combination	AND ● / OR ○	AND ● / OR ○

This table can be extended by additional compound classes if desired

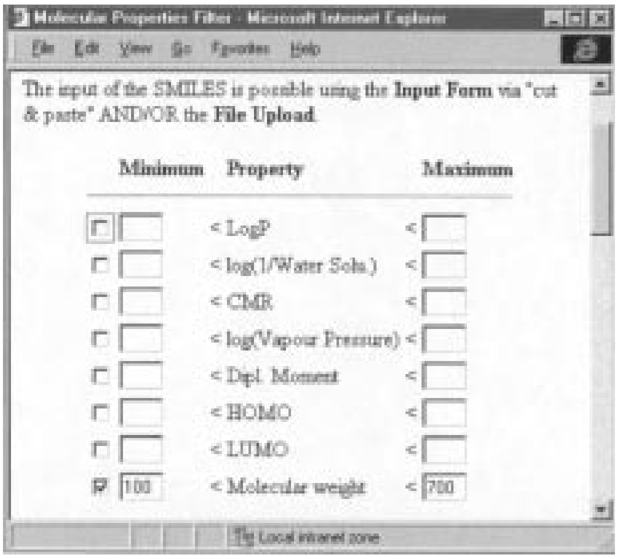
☑ remove compounds with unusual elements
not H, C, N, O, S, P, F, Cl, Br, I, Li, Na, K, Ca, Mg

Fig. 10. Web interface of the molecular structure filter.

Diversity and filter tools

The Diversity and Filter tools have a simple web interface for easy and intuitive handling. The tools are bundled into one single module which contains submodules for diversity-based selection, the molecular property filter, the molecular structure filter and the hitlist viewer. Several input formats can be used: SMILES strings, hitlists of compound numbers or structure files in SMILES and SD formats. A hitlist can be created by the diversity selection tools, the property or structure filter or WinMerlin, our in-house chemical datamining program. The tools are designed to handle data sets of up to 50,000 molecules.

The *molecular structure filter* (Fig. 10) removes compounds which fulfill or do not fulfill certain filter conditions. A filter condition is the existence or absence of a



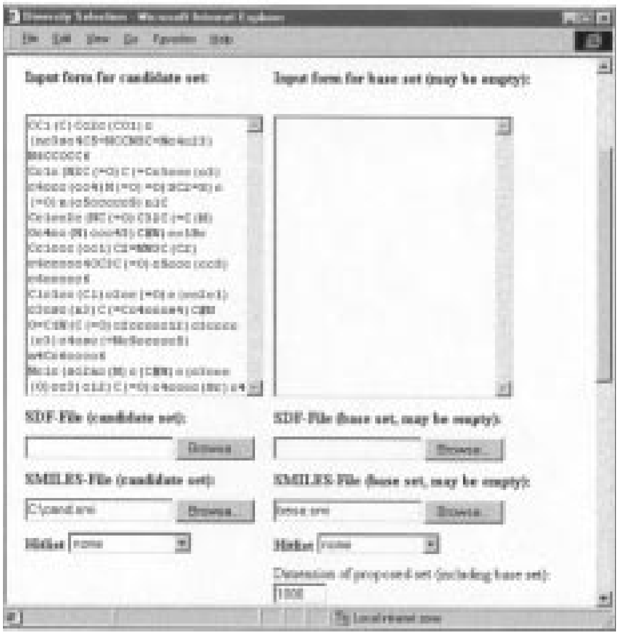
The input of the SMILES is possible using the **Input Form** via "cut & paste" AND/OR the **File Upload**.

	Minimum	Property		Maximum
□		< LogP	<	
□		< log(1/Water Sols.)	<	
□		< CMR	<	
□		< log(Vapour Pressure)	<	
□		< Dipl. Moment	<	
□		< HOMO	<	
□		< LUMO	<	
☑	100	< Molecular weight	<	700

Fig. 11. Web interface of the molecular property filter.

functional group or a chemical element. Several conditions can be combined by the logical AND and OR operators or by applying different filters sequentially.

The *molecular property filter* (Fig. 11) excludes all compounds from the given set of molecules which do not have a given range of properties. The properties logP, water solubility, vapor pressure, molar refractivity, dipole moment, molecular weight, HOMO and LUMO are stored in a database for fast retrieval. The filter program calls the property calculator (see below) if the properties are not already stored in the database.



Input form for candidate set:

```
[illegible]
```

Input form for base set (may be empty):

SDF-File (candidate set):

SDF-File (base set, may be empty):

SMILES-File (candidate set): C:\pand.smi

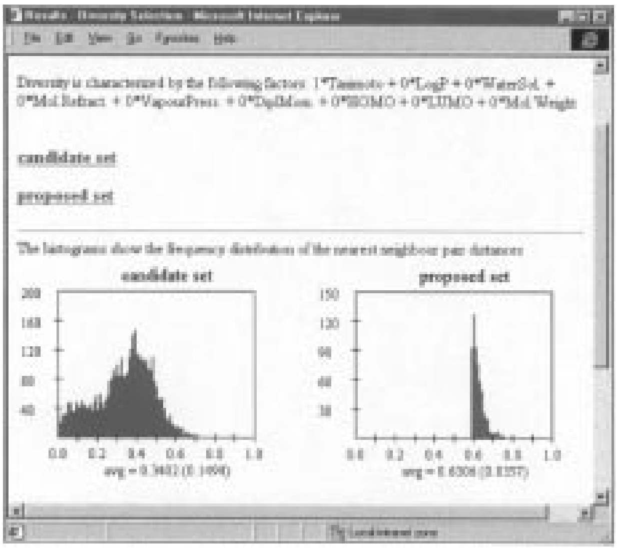
SMILES-File (base set, may be empty): base.smi

Hitlist none

Hitlist none

Dimension of proposed set (including base set): 1000

Fig. 12. Web interface of the diversity-based selection program.



Diversity is characterized by the following factors: 1*Tanimoto + 0*LogP + 0*WaterSol. + 0*Mol.Refract. + 0*VapourPress. + 0*DiplMom. + 0*HOMO + 0*LUMO + 0*Mol.Weight

candidate set

proposed set

The histograms show the frequency distribution of the nearest neighbour pair distances

Fig. 13. The results page of the diversity-based selection program, containing nearest neighbor histograms for the candidate set and the proposed set.

The *diversity module* (Fig. 12) allows the selection of a diverse subset of molecules from a given candidate set. It is also possible to extend a given set (base set, *e.g.*, the set of molecules which were already screened) with molecules from another set of molecules (*e.g.*, the set which is formed from offers of third party suppliers). In the latter case, the set selected is diverse within itself and also diverse to the molecules in the reference set.

The distance definition, on which the diversity selection is based, can be modified interactively. The structural distance can be defined by the Tanimoto coefficient or by differences in logP, water solubility, vapor pressure, molar refractivity, dipole moment, molecular weight, HOMO and LUMO. It is also possible to apply different weights to the properties. Before applying the weighting factors, the distances defined by different properties are scaled to the range [0.1]. The Hamming metric is then applied to combine the various properties.

The results page of a diversity selection run is shown in Figure 13. The selected set is available via the link "proposed set". To get an impression of the diversity of the selected set, nearest neighbor distances can be drawn. A peak at large distances indicates that the selected set is diverse. A sharp peak (low standard deviation) shows that nearest neighbor distances for all nearest neighbor pairs are similar. This means that the chemical space is covered evenly and that there are no accumulations and no holes. One indication of high diversity is a high average nearest neighbor distance.

The *hitlist viewer* (Fig. 14) is used to inspect more closely the hitlist created by the filter and diversity modules. It allows to rename and delete hitlists. In addition, it can be used to export the hitlist to the client computer, from which they can be loaded into other programs such as WinMerlin, and to visualize the compounds and their properties.

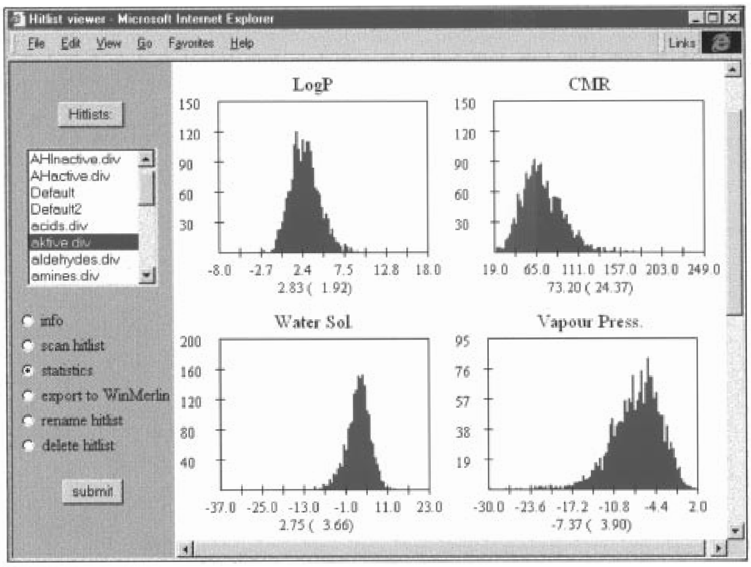


Fig. 14. The hitlist viewer is used to visualize property distributions of a compound subset. The range, average and standard deviation of property values are shown.

Property calculation tool

Physicochemical properties of molecules under consideration for screening are calculated with a web-based property calculator. This program enables easy interactive calculations of properties for single molecules, as well as batch calculations for whole sets of molecules. Calculated properties include hydrophobic and electronic properties.

Hydrophobic properties determine the ability of a molecule to be transported in the environment and in an organism, to interact with biological membranes and to bind to a receptor by hydrophobic forces. The hydrophobic properties which are calculated by our system are logP (logarithm of the octanol-water partition coefficient), MR (molar refractivity), log(1/WS) (water solubility) and log(VP) (vapor pressure). They are calculated with in-house programs based on published theories (30, 31, 48).

Electronic properties characterize the electronic distribution within the molecule. They account for the ability of a molecule to react, as well as for the electronic interaction with a receptor. Electronic properties are calculated by the AM1 (49) semiempirical method. Available electronic properties are dipole moment, and the energies of the HOMO (highest occupied) and LUMO (lowest unoccupied molecular orbital).

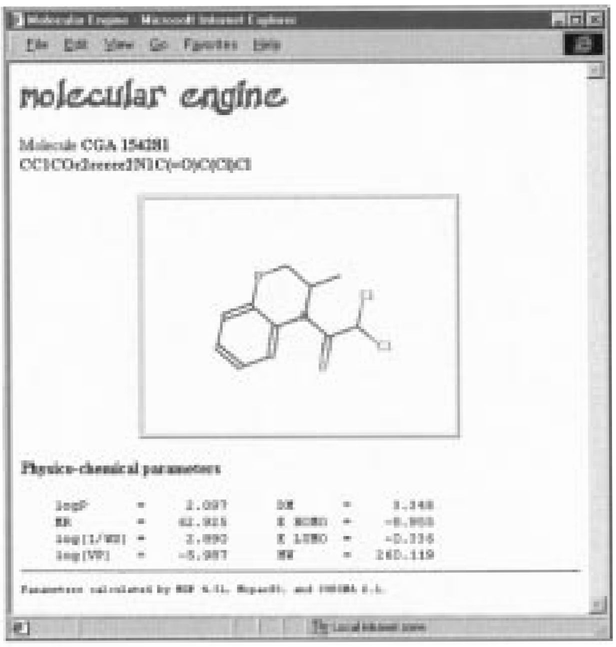
When the property calculator is used interactively, molecules are simply entered into the system in the form of their company test number as a SMILES string, or they are drawn with the help of a molecular editor written in Java. Once a job is submitted, a relatively complex chain of processes is started. Various programs are launched, including the CORINA 3D builder (50) which creates a 3D molecular geometry, the Mopac93 package (32) which calculates electronic parameters, and the in-house programs which calculate hydrophobic properties. In spite of this complex processing, response time is short and the results are delivered within 4-5 seconds (Fig. 15). This interactive module is part of our Novartis web-based molecular modelling system (51). The property calculator may also be called directly (without the graphic interface) by referencing the http address of the cgi script, with a SMILES supplied as parameter. In this way, it is possible to calculate data for a large number of molecules in "batch" mode. By using this technique, data for the more than 800,000 molecules, which can be used in the diversity selector, were generated.

Conclusions and outlook

We have developed a number of compound selection methods based on structural features and computed physicochemical parameters. Structure-based diversity selection methods are used to ensure that compounds which are purchased from external suppliers of screening substances as input into our HTS program are sufficiently novel. To enrich screening collections with compounds which have a higher probability of being active, we find that structure-based selection methods perform better than methods based on physicochemical parameters.

The selection tools which we have developed can be used, in principle, by all chemists at Novartis Crop Protection. In practice, they are used by specialists who are involved in compound acquisition and screening logistics. Activity optimization by genetic algorithms is still in an experimental phase and an area for specialist work.

There are several goals which we are still actively pursuing. Foremost is the development of selection methods which combine structural diversity and favorable molecular properties, and which should ideally lead to compound


Molecular Engine - Microsoft Internet Explorer
File Edit View Go Favorites Help
Local intranet zone

molecular engine

Molecule CGA 154281
CC1COc2ccccc2N1C(=O)C(Cl)Cl

Physico-chemical parameters

```
    logP      =     2.097    DM      =     3.348
    MR        =    62.925    E HOMO  =    -8.955
    log(1/WS) =     2.890    E LUMO  =    -0.336
    log(VP)   =    -5.987    MW      =   260.119
```

```
Parameters calculated by MOE 4.11, Mopac93, and CORINA 1.1.
```

Fig. 15. Web-based calculation of molecular physicochemical parameters.

subsets rich in novel and active compounds. Furthermore, a careful optimization of fragments should improve the performance of structure-based activity prediction methods.

The most important area, however, continues to be the empirical validation of diversity selection methods and their comparison with the chemists' traditional method of inspection and intuition. Running simulated compound selection experiments on historical databases is a useful device to develop parameter schemes and selection algorithms, but does not allow to compare the theoretical and manual selection methods. Moreover, it does not address difficult practical issues such as how to integrate these methods into a large-scale and highly standardized screening effort. The proof that diversity methods are not only plausible and theoretically interesting, but also useful, will have to come from real screening programs and will require years of experience.

Acknowledgements

We would like to thank Heinz Binder and the HTS Logistics Department, and especially Dr. J.-C. Gehret, for the good collaboration in building the database with third-party contributions.

References

1. Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A., Gordon, E.M. *Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries.* J Med Chem 1994, 37: 1233-51.

2. I.M. Chaiken, K.D. Janda (Eds.). Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery. American Chemical Society: Washington 1996.

3. Felder, E., Poppinger, D. *Combinatorial compound libraries for enhanced drug discovery approaches.* Adv Drug Res 1997, 30: 111-99.

4. Willett, P. *Computational tools for the analysis of molecular diversity.* In: Perspectives in Drug Discovery and Design, Vol. 7/8: Computational Methods for the Analysis of Molecular Diversity. P. Willett (Ed.). Kluwer Academic Publishers: Dordrecht 1997, 1-11.

5. Weininger, D. *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* J Chem Inf Comput Sci 1988, 28: 31-6.

6. ChemBridge Corporation website: http://www.chembridge.com/.

7. SPECS and BioSPECS: http://www.specs.net/.

8. Sigma-Aldrich Library of Rare Chemicals: http://www.sigma.sial.com/aldrich/rare_chemicals/.

9. Lynch, M.F., Barnard, J.M., Welford, S.M. *Generic structure storage and retrieval.* J Chem Inf Comput Sci 1985, 25: 264-70.

10. G.M. Maggiora, M.A. Johnson (Eds.). Concepts and Applications of Molecular Similarity. Wiley: New York 1990.

11. Good, A.C., Mason, J.S. *Three-dimensional structure database searches.* In: Reviews in Computational Chemistry, Vol. 7. K.B. Lipkowitz, D.B. Boyd (Eds.). VCH: New York 1996, 67-117.

12. Cramer, R.D., Clark, R.D., Patterson, D.E., Ferguson, A.M. *Bioisosterism as a molecular diversity descriptor: Steric fields of single "topomeric" conformers.* J Med Chem 1996, 39: 3060-9.

13. Kearsley, S.K., Sallamack, S., Fluder, E.M., Andose, J.D., Mosley, R.T., Sheridan, R.P. *Chemical similarity using physiochemical property descriptors.* J Chem Inf Comput Sci 1996, 36: 118-27.

14. Martin, Y.C., Bures, M.G., Brown, R.D. *Validated descriptors for diversity measurements and optimization.* Pharm Pharmacol Commun 1998, 4: 147-52.

15. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Weinberger, L.E. *Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors.* J Med Chem 1996, 39: 3049-59.

16. James, C.A., Weininger, D. *Daylight Software Manual Version 4.42.* Daylight Chemical Information Systems, Inc.: Irvine 1996. See also: http://www.daylight.com/dayhtml/doc.

17. *MACCS-II.* Molecular Design, Ltd.: San Leandro, CA, USA.

18. Tripos Associates, 1699 S. Hanley Road, Suite 303, St. Louis, MO 63144, USA.

19. Willett, P., Winterman, V. *A comparison of some measures for the determination of inter-molecular structural similarity: Measures of intermolecular structural similarity.* Quant Struct-Act Relatsh 1986, 5: 18-25.

20. Davies, K., Briant, C. *Combinatorial chemistry library design using pharmacophore diversity.* Network Science (http//www.awod.com), July 1995.

21. Lajiness, M.S., Johnson, M.A., Maggiora, G.M. *Implementing drug screening programs by using molecular similarity methods.* In: QSAR: Quantitative Structure-Activity Relationships in Drug

Design. J.L. Fauchere (Ed.). Alan Liss, Inc.: New York 1989, 173-6.

22. P. Willett (Ed.). Similarity and Clustering in Chemical Information Systems. Research Studies Press: Letchworth 1987.

23. H. Martens, T. Naes (Eds.). Multivariate Calibration. Wiley: New York 1989.

24. Martin, E.J., Blaney, J.M., Siani, M.A., Spelimeyer, D.C., Wong, A.K., Moos W.H. *Measuring diversity: Experimental design of combinatorial libraries for drug discovery.* J Med Chem 1995, 38: 1431-6.

25. Gillet, V.J., Willett, P., Bradshaw, J. *The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries.* J Chem Inf Comput Sci 1997, 37: 731-40.

26. Gobbi, A., Poppinger, D., Rohde, B. *Developing an inhouse system to support combinatorial chemistry.* In: Perspectives in Drug Discovery and Design, Vol. 7/8: Computational Methods for the Analysis of Molecular Diversity. P. Willett (Ed.). Kluwer Academic Publishers: Dordrecht 1997, 131-58.

27. Carhart, R.E., Smith, D.H., Ventkatataraghavan, R. *Atom pairs as molecular features in structure-activity studies: Definition and applications.* J Chem Inf Comput Sci 1985, 25: 64-73.

28. Moreau, G. *Use of similarity analysis to reduce large molecular libraries (inhouse, combinatorial or commercial) to smaller sets of representative molecules.* Synthetic Chemical Libraries in Drug Discovery (Oct 30-31, London) 1995.

29. Dunbar, J.B. *Cluster-based selection.* In: Perspectives in Drug Discovery and Design, Vol. 7/8: Computational Methods for the Analysis of Molecular Diversity. P. Willett (Ed.). Kluwer Academic Publishers: Dordrecht 1997, 51-63.

30. Viswanadhan, V.N., Revankar, G.R., Robins, R.K. *Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships.* J Chem Inf Comput Sci 1989, 29: 163-72.

31. Wakita, K., Yoshimoto, M., Miyamato, S., Watanabe, H. *A method for calculation of the aqueous solubility of organic compounds by using new fragment solubility constants.* Chem Pharm Bull 1986, 34: 4663-81.

32. Stewart, J.J.P. *MOPAC 93*. Fujitsu Ltd., Tokyo, Japan, 1993, available from Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, USA.

33. Cramer, R.D., Patterson, D.E., Bunce, J.D. *Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins.* J Am Chem Soc 1988, 110: 5959-67.

34. Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J., Humblet, C. *Stigmata: An algorithm to determine structural commonalities in diverse datasets.* J Chem Inf Comput Sci 1996, 36: 862-71.

35. Klopman, G. *Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules.* J Am Chem Soc 1984, 106: 7315-21.

36. Enslein, K., Craig, P.N. *A toxicity estimation model.* J Environ Pathol Toxicol 1978, 2: 115-21.

37. Benigni, R. *The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: Definitive results.* Mutat Res 1997, 387: 35-45.

38. Cramer, R.D., Bunce, J.D., Patterson, D.E., Frank, I.E. *Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies.* Quant Struct-Act Relatsh 1988, 7: 18-25. Erratum 1988, 7: 91.

39. Weber, L., Wallbaum, S., Broger, C., Gubernator, K. *Optimization of the biological activity of combinatorial compound libraries by a genetic algorithm.* Angew Chem 1995, 107: 2452-4.

40. Singh, J., Ator, M.A., Jaeger, E.P. et al. *Application of genetic algorithms to combinatorial synthesis: A computational approach to lead identification and lead optimization.* J Am Chem Soc 1995, 118: 1669-76.

41. Myers, P.L., Greene, J.W., Saunders, J., Teig, S.L. *Rapid, reliable drug discovery.* Today's Chem Work 1997, 6: 46-48, 51, 53.

42. Gobbi, A., Poppinger, D. *New leads by selective screening of compounds from large databases.* 213th ACS Natl Meet (Apr 13-17, San Francisco) 1997, Abst CINF-067.

43. Michalewicz, Z. (Ed.). Genetic Algorithm + Data Structures = Evolution Programs. Springer Verlag: Berlin 1996.

44. Gobbi, A., Poppinger, D. *Genetic optimization of combinatorial libraries.* Biotechnol Bioeng 1998, 61: 47-53.

45. Bremser, W. *HOSE - A novel substructure code.* Anal Chim Acta 1978, 103: 355-65.

46. Brown, R.D., Martin, Y.C. *The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding.* J Chem Inf Comput Sci 1997, 37: 1-9.

47. Weinstein, J.N., Myers, T.G., O'Connor, P.M. et al. *An information-intensive approach to the molecular pharmacology of cancer.* Science 1997, 275: 343-49. See also NCI database, http://epnws1.ncifcrf.gov:2345/.

48. Lyman, W.J., Reehl, W.F., Rosenblath, D.H.R. (Eds.) Handbook of Chemical Property Estimation Methods. American Chemical Society: Washington 1990.

49. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F., Stewart, J.J.P. *AM1: A new general purpose quantum mechanical molecular model.* J Am Chem Soc 1985, 107: 3902-9.

50. Sadowski, J., Gasteiger, J. *From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders.* Chem Rev 1993, 93: 2567-81.

51. Ertl, P., Jacob, O. *WWW-based chemical information system.* THEOCHEM 1997, 419: 113-20. See also P. Ertl, http://www.elsevier.com/homepage/saa/eccc3/paper6.