

[F0008]

Finding Biological Active Compounds in Large Databases

[Alberto Gobbi](#)^{*}, [D. Poppinger](#), [B. Rohde](#)

Novartis AG, Postfach, CH-4002 Basel, Switzerland

E-mail: (Alberto Gobbi) Alberto.Gobbi@cp.novartis.com,
(Dieter.Poppinger) Dieter.Poppinger@cp.novartis.com, and (B. Rohde)
Bernhard.Rohde@pharma.novartis.com

Received: 27 August 1997 / Uploaded: 28 August 1997

Abstract

A method is proposed, and shown to be able to find active compounds in a Large Database by iterative selection/screening cycles. The method is derived from genetic algorithms. Starting from an initial small parent population the next compounds to screen are selected by similarity search based on the structural features of two high rating parents. The performance of the method is demonstrated using 20000 compounds with biological data from the NCI¹ database.

Table of Contents:

- [Finding Biological Active Compounds in Large Databases](#)
- [Lead Finding As an Optimization Process](#)
- [Genetic Algorithm](#)
- [Modified Genetic Algorithm for Lead Finding](#)

- [A Few Words on Similarity](#)
 - [Validation](#)
 - [Results](#)
 - [Conclusions](#)
-

Finding Biologically Active Compounds in Large Databases

Given a large database with compounds that are available in-house or through an external supplier many agrochemical and pharmaceutical companies face the problem of how to select compounds for screening. The major goal is to find the active compounds as fast as possible.

There are several approaches one could imagine to find active compounds in large databases. Random screening is the conceptually most easy and therefore many pharmaceutical and agrochemical companies have set up large random screening programs². The opposite approach is given by sophisticated modeling techniques and 3D database searching based on rational pharmacophore models³.

Much research is now ongoing in the field of diversity selection, where the aim is not directly to find biological active compounds but to scan the given chemical space using as few as possible compounds. Since the aim of diversity selection is to find diverse compounds diversity selection will not yield more active compounds but it will yield more diverse active compounds. Statistically, diversity selection will not give a higher number of hits as shown in figure 1.

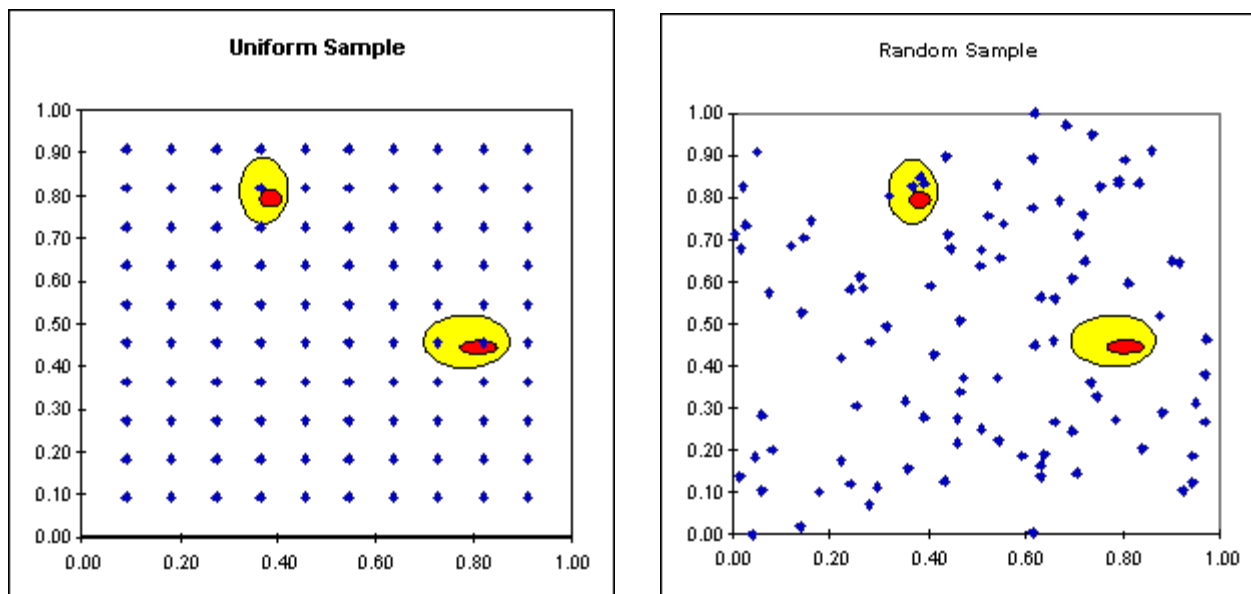


Figure 1. Schematic depiction of selection by diversity design, where compounds are chosen to cover the space evenly (left) and randomly (right). The red spots mark areas where highly active compounds may be found. The yellow spots mark areas where moderately active compounds may be found.

On the other hand diversity selection offers two advantages over random selection:

- Since diversity selection cover the chemical space evenly, two sets selected by diversity out of the same database will cover the same classes of compounds. On the other hand different random selections might differ widely in the number and kind of chemical classes covered. Even worse, if large clusters are present, a random selection will tend to be biased towards selecting compounds out of these clusters.
- By using diversity techniques the chances are increased that active compounds from all or at least most activity classes may be found.

Lead Finding as an Optimization Process

A closer look at the lead finding process in agrochemical and pharmaceutical companies reveals that it can be regarded as an iterative optimization process (figure 2). Starting from some weakly

active compound modifications are made to the chemical structure to improve activity. These modifications are usually driven by intuition, QSAR approaches or molecular modeling. The newly synthesized compounds are screened, and then might be used themselves again as the basis for further modifications until sufficient activity is found. This process has to be iterative since the exact relation between the chemical structure and activity is unknown.

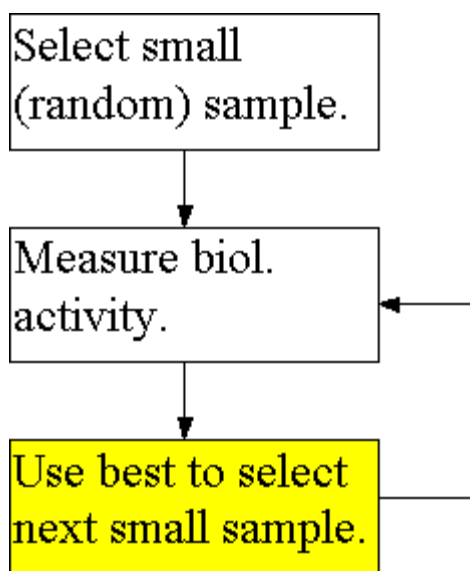


Figure 2. Iterative optimization of activity.

However, optimization problems do not only exist in chemistry, but are common in any scientific or technical field. Therefore many computational algorithms are available to solve optimization problems⁴. The optimization problem is usually formulated as follows:

- Given a function $f(x)$ find those x where $f(x)$ has a maximum while x is in a given range.

The lead finding problem may be formulated in analogy:

- Given a biological screen f for compounds x , find the compounds with highest activity in a given database of available compounds.

Genetic algorithms have proven to be able to solve hard optimization

problems , therefore we have implemented an optimization algorithm derived from genetic algorithms⁶.

Genetic Algorithm

Genetic algorithms try to emulate the way in which nature optimizes its species through evolution. Given a set of individuals, those with higher fitness are allowed to mate more often and therefore to produce more offspring. To produce offspring, the chromosomes of two individuals are combined in the crossover step. As in nature, in addition to crossover the chromosome of the offspring may be modified by mutation. The new chromosome defines a new member for the next iteration. The computational deployment of genetic algorithms is strongly dependent on the possibility to create something which is analogous to biological chromosomes. In standard genetic algorithms, bit strings are used as chromosomes. If, for example, the highest value of the function:

$$f(x) = x * \sin(10 \pi * x) + 1$$

in the interval -1 - 2 is to be found, x has to be represented as a bit string. This can be accomplished by mapping the desired range (-1 - 2) onto a 32 bit integer value which ranges from 0 to 4294967296. The value of x may be derived from xi by:

$$x = -1 + \frac{xi}{4294967292} * 3$$

which yields values between -1 and 2 as required.

A standard genetic algorithm would use a procedure like the following:

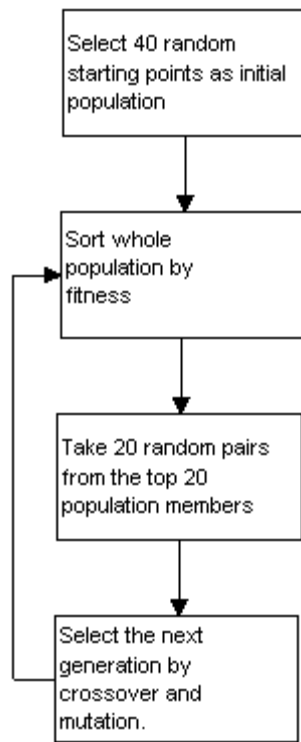


Fig 3. Standard genetic algorithm

1. Select an initial random population of e.g. 40 members:

	bin(xi)	xi	x	f(x)
1	000000101011101110101111100110	22927334	-0.98398	0.90954
2	110010100001110011111100010111	847724311	-0.407871	0.49450
3	111010101101100010101110001001	2058759049	0.43801	1.10418
4			

2. Select n random pairs of parents from the n members of the population with highest fitness (largest $f(x)$). E.g. number 1 and 3.
3. Crossover the chromosomes (x_i) of the selected pairs at a random position. E.g.

110010100001110011111100010111 and 111010101101100010101110001001
 yield 110010100001110011111110001001

4. Allow for a limited number of random mutations. E.g.
5. 110010100001110011111110001001 yields
110000100001110011111010001001
6. Calculate the fitness $f(x)$ for the new children and add them to the population.
7. Repeat steps 2 to 5 until the maximum value of $f(x)$ is found.

Genetic algorithms work because the chromosomes of high-rating individuals have higher chance to survive and because the features of

high rating individuals are combined to give even better individuals in the next generation. Mutations allow for the exploration of unknown regions in the search space.

Modified Genetic Algorithm for Lead Finding

The procedure described above should also lend itself to lead finding. In much the same way as in conventional lead optimization, starting from a small number of screened compounds, new ones should be screened in order to find more active compounds. The key problem in using a genetic algorithm to optimize the outcome of a biological screen is the bit string representation of a chemical structure. A string containing the flask number of reagents for a virtual combinatorial library has been used as chromosome for a genetic algorithm by Weber et. al.⁷. The genetic algorithm converged quickly to find active compounds.

On the other hand most chemical information systems already contain a bit string coding for structures (Fig 4). These were introduced to speed up substructure searches and to allow for similarity searching². For our work we have chosen to use fingerprints as created with the [daylight toolkits](#). A "1" within a fingerprint codes for the presence of one or more structural features while a "0" shows that the corresponding fragments are absent.

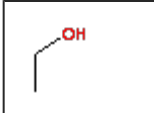
	CN	COC	CC	CO	CCC	C	CCN	O	N	CCO
	0	0	1	1	0	1	0	1	0	1

Fig. 4. Fingerprint of ethanol (simplified).

This molecular fingerprints are an ideal representation which can be used in the crossover step of a genetic algorithm (Fig 5).

	COC		CCC		CCN		CCO	
	CN	CC	CO	C	O	N		
CCO	0	0	1	1	0	1	0	1
CCN	1	0	1	0	0	1	1	0
Crossover	1	0	1	0	0	1	0	1

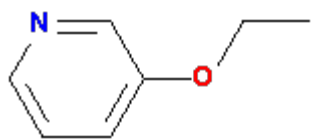
Fig 5. Fingerprint crossover.

The crossover fingerprint encodes features from both parents. However, it will usually not correspond directly to a chemical structure from the database. Rather, the database molecule which corresponds most closely to the combination of both parents can be found by a simple similarity search through the database.

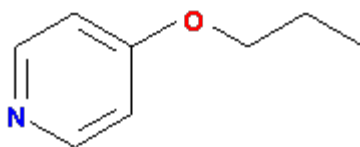
A Few Words on Similarity

One of the most often used similarity measures based on fingerprints is the Tanimoto coefficient²:

$$Tanimoto(1,2) = \frac{\text{Bits in common}(1,2)}{\text{Total Bits Set}(1,2)}$$



1



2

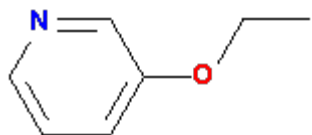
0.79

$$Tanimoto(1,2) = 0.79$$

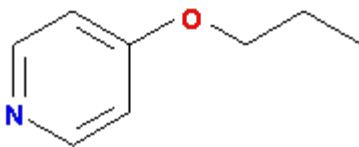
The Tanimoto coefficient varies between 0, both structures have no structural feature in common, and 1 both structure contain exactly the same structural features.

In addition we have implemented an substructure similarity measure:

$$\text{Substructure Sim}(1,2) = \frac{\text{Bits in common}(1,2)}{\text{Bits Set}(1)}$$



1



2

Substructure Sim(1,2)=0.85

The substructure similarity also varies between 0 and 1. But the substructure similarity measures to which extent 1 is a substructure from 2, yielding 1.0 if 1 is an exact substructure from 2.

Validation

To validate the method we have simulated the iterative screening/selection cycle using compounds and data from the NCI¹ database. The non-small cell lung cancer results using the A549/ATTC cell line, have been chosen because of the large number of screened compounds. The GI50 values given in the database are corrected IC50 values. As can be seen in figure 6 more than half of the compounds measured have an $-\log(\text{GI50})$ value of exactly 4.0. Only 4 compounds have $-\log(\text{GI50}) > 11.75$.

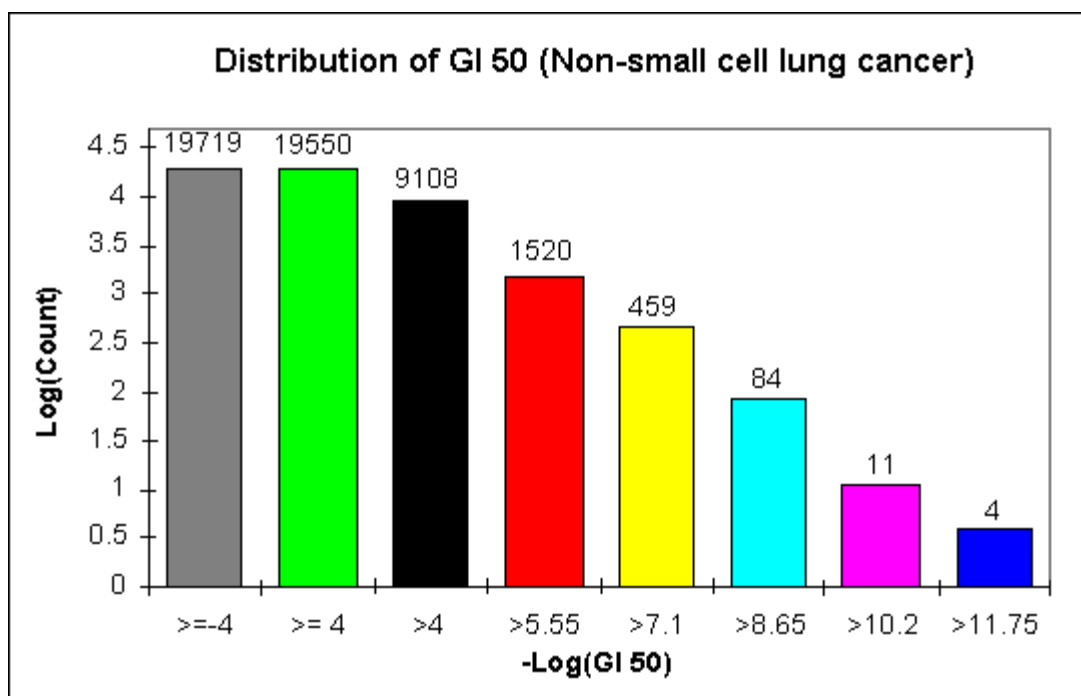


Figure 6. Distribution of the negative logarithm of the GI50 value of the inhibition of non small cell lung cancer for 19719 compounds, as given in the NCI database. The values range from -4 (inactive) to 13.0 (active). The range between 4 and 13 was divided in 6 equally spaced groups. More than half of the compounds have an activity of 4.0.

The aim of the optimization should be to find the compounds of the active groups in as few as possible steps. The procedure was as follows:

- Twenty compounds were selected randomly out of the set of compounds with $\log(\text{GI}50) = 4.0$ and added to the set of screened compounds with known activity. The following steps were then iterated 100 times:
 - Take 20 most active compounds out of the set with known activity to build the parent set.
 - Add 2 random compounds from the database to the parent set (mutation). This allows for the exploration of more diverse regions of the chemical space.
 - Create 20 crossover fingerprints by randomly selecting 20 pairs from the parent set.
 - For each crossover fingerprint lookup the most similar compound in the set of unscreened compounds and add it to the set of screened compounds.
 - If a compound was included more than 10 times in a parent set do not use it again.

Results

The results using the substructure similarity are given in figure 7. Figure 7 shows a typical run. Several runs might differ by the initial seed to the pseudo random number generator, or by the choice of the first parent generation. However the performance in all our experiments has been comparable to that shown in figure 7. Using a random screening approach one would expect that all lines collapse onto the green line which shows the percentage of compounds screened. As can be seen the algorithm performs much better than random screening. After screening only about 5% of the database all compounds from the highest active set were found and about 70% of the compounds with $\log(\text{GI50}) > 10.2$.

Inspection of the structures from the most active set given on the left shows that they do not fall into one single class. Moreover they are quite different from each other. They may be classified into two pairs one containing two macrocyclic compounds. Thus the method seems to be able to find different maxima in the activity space.

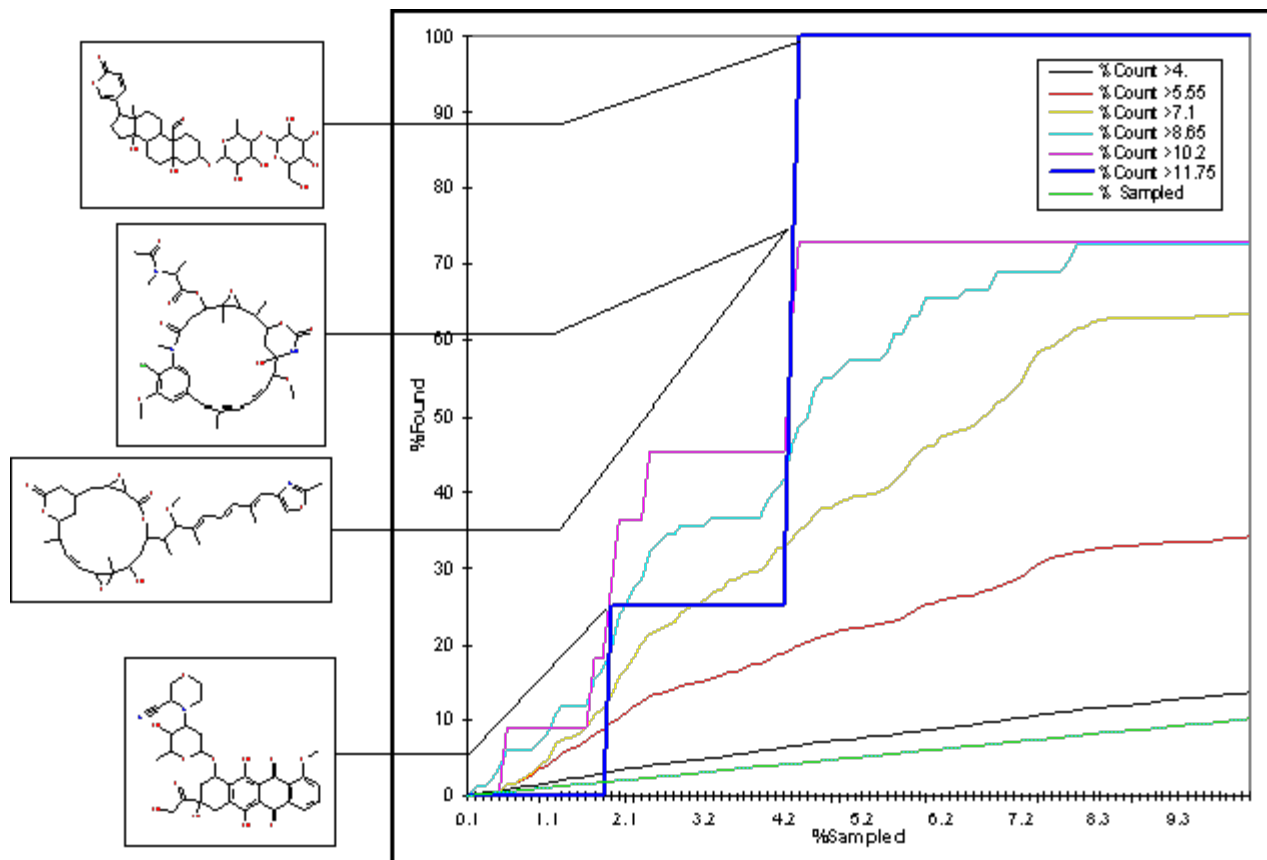


Figure 7. Simulated lead finding using the modified genetic algorithm and substructure similarity. The percentage of compounds found from most active sets of compounds in figure 6 is plotted against the percentage of compounds screened. Using a random screening approach one would expect that all lines collapse onto the green line which shows the percentage of compounds screened.

The results using the Tanimoto similarity are given in figure 8. Although the results are still two to three times better than random screening, the substructure similarity is much better. This might be understood if one assumes that the presence of some structural features is essential for activity. Additional structural features in most cases will not harm the activity unless they are interfering with the receptor. Therefore additional structural feature should not be weighted as much as absent structural features when looking for active compounds. This is exactly what happens when the substructure similarity is used.

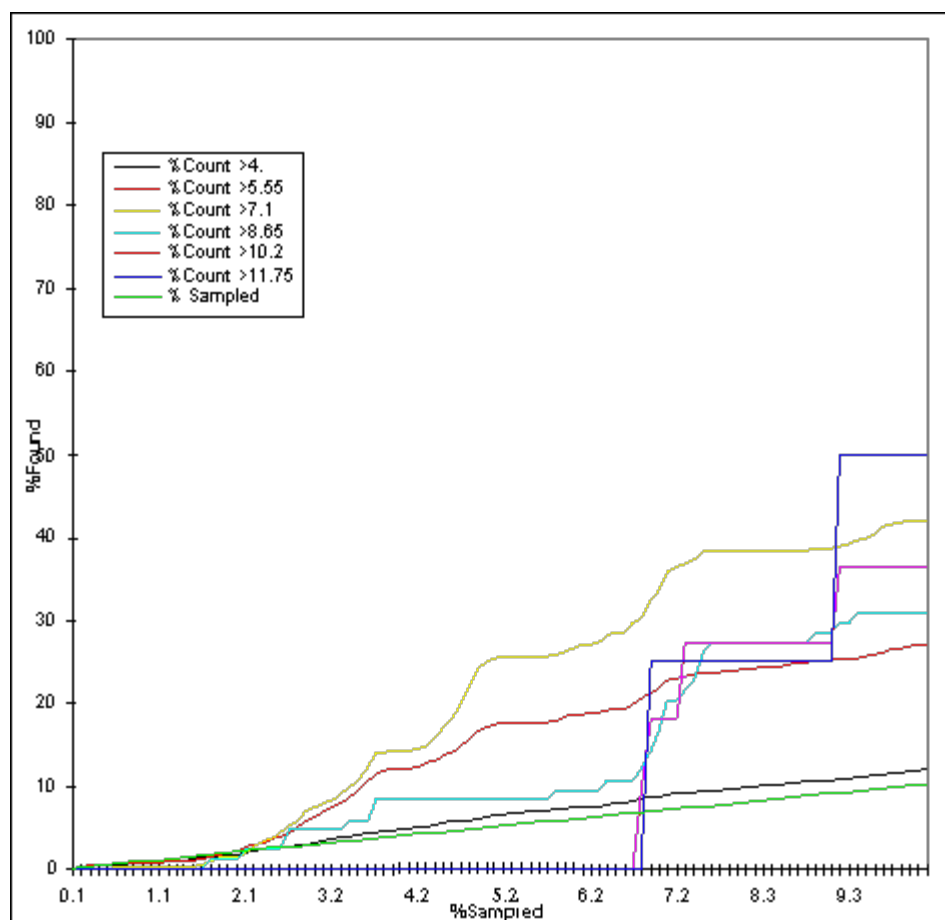


Figure 8. Simulated lead finding using the modified genetic algorithm and Tanimoto similarity.

Conclusions

By using optimization techniques based on structural similarity it is possible to find active compounds out of large databases, while screening only a small fraction of the complete database. An algorithm derived from genetic algorithms was proposed and its performance analyzed by simulating the selection/screening iterations using data from the National Cancer Institute¹. After screening only about 5% of the compounds all compounds out of the most active group were found. While the crossover of two fingerprints combines features from two previously active compounds, the addition of mutations assures that unknown parts of the chemical space are explored. Therefore the proposed algorithm can find active compounds from several activity groups.

- [1] NCI database: <http://epnws1.ncifcrf.gov:2345/>,
N. Weinstein, T. G. Myers, P. M. O`Connor, S. H. Friend, A. J. Fornance, 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
- [2] John P. Devlin, <http://www.awod.com/netsci/Science/Combichem/feature14.html>
- [3] G. M. Downs, P. Willet in *Reviews in Computational Chemistry*, K. Lipkowitz, D. B. Royd (Editors), VCH
Publishers Inc. New York 1996, 1-66
- [4] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C, The Art of
Scientific Computing*, Second Edition, Cambridge University Press 1992
- [5] Z. Michalewicz, *Genetic + Data Structures = Evolution Programs*, Springer Verlag Berlin Heidelberg New
York 1996
- [6] A. Gobbi, D. Poppinger, B. Rohde *J. Chem. Inf. Sci.* submitted for publication.
- [7] L. Weber, S. Wallbaum, C. Broger, K. Gubernator *Angew. Chem.* 1995, 107, 2452
-

Comments

During 1-30 September 1997, all comments on this poster should be sent by e-mail to
ecsoc@listserv.arizona.edu with **F0008** as the message subject of your e-mail. After the conference, please
send all the comments and reprints requests to the author(s).
