Evolutionary combinatorial chemistry: application of genetic algorithms

Lutz Weber

Evolutionary chemistry combines the evaluation of molecular properties and synthesis of novel compounds in a feedback loop to arrive at molecules with the desired properties. Inspired by natural evolutionary processes, combinatorial chemistry in combination with mathematical optimization methods and biological testing provides new approaches to drug discovery. Genetic algorithms have been applied with success in the design and automated synthesis of combinatorial compound libraries.

The idea of evolutionary processes in chemistry is as old as chemistry itself. Wöhler's successful experiments to generate oxalic acid from dicyane in 1824 and urea from ammonium cyanate in 1828 were understood at that time as the evolution of inorganic matter into organic molecules, which are the basis of life. The formation of new molecules in living systems is today understood in terms of darwinian evolution, in which new molecules evolve on the basis of feedback on their properties and their contribution to the performance of the organism. A striking example is the evolution of antibodies that are assembled from a 'combinatorial DNA library' of the variable V, D and J gene sequences in the primary immune response. Antibodies that bind to an antigen are expressed on B lymphocytes, and subsequent somatic hypermutation yields more 'fit' antibodies with better binding affinities. The invention of novel drugs by medicinal chemists can be understood as a similar evolutionary process; thus, new, better molecules are synthesized in a feedback loop based on the various properties of previous molecules.

Large compound libraries

The recent introduction of novel high-throughput methods in drug discovery, such as combinatorial chemistry and screening for biological and physicochemical properties, has allowed chemists to generate and test many molecules virtually simultaneously. We are thus in a similar position to nature: that of being able to synthesize many molecules out of the billions possible from a combinatorial library using a rather limited number of starting materials, but not knowing a priori which molecule out of all possible molecules will satisfy our criteria for success.

Most pharmaceutical companies have begun to realize, for reasons of cost and time, that there is no point in producing and testing very large libraries of a given compound class (e.g. all 64 million hexapeptides). Screening large combinatorial libraries might also provide redundant information when structurally similar molecules are tested. False positive and false negative screening hits further increase the data 'noise'. However, does the quantitative increase in the possibilities to generate and test large numbers of compounds also allow for a new dimension in the drug discovery process? An answer to this question could be the novel approach to evolutionary chemistry that tries to connect the selection and synthesis of biologically active compounds with mathematical optimization methods. Heuristic algorithms such as the genetic algorithm or neuronal networks mimic darwinian evolution and do not require a priori

Lutz Weber, Morphochem AG, Am Klopferspitz 19, 82152 Martinsried, Germany. tel: +49 89 74 00 70 0; fax: +49 89 74 00 70 80, e-mail: lutz.weber@morphochem.com

  1359-6446/98/$19.00. PII: S1359-6446(98)01219-7

knowledge of SARs. These combinatorial optimization methods[1] have proved useful in solving multidimensional problems and are now being used with success in various areas of combinatorial chemistry. Thus, evolutionary chemistry could aid in the selection of information-rich subsets of available compound libraries or in designing screening libraries and new compounds to be synthesized, thereby adding a new dimension to combinatorial chemistry.

Genetic algorithms

Many mathematical algorithms have been developed to select optimal combinations from a pool of combinatorial possibilities. One of the most appealing is the genetic algorithm (GA), a machine learning technique originated by Holland[2] and Rechenberg[3] to solve combinatorial optimization problems. A GA starts with a set of different entities (e.g. molecules) encoded by genomes (the population). A ranking of these parent genomes is then performed according to the measured or calculated fitness of the corresponding entities. New genomes are then generated for the next generation in the selection step from this ranked list by the GA functions: replication, mutation and crossover (Fig. 1). Replication regenerates an equivalent genome; mutation sets a bit (gene) in the parent genome to a different value to obtain a new genome; and crossover takes two or more genomes to build new ones by mixing them according to various rules. After the new set of genomes (children) has been generated, they are translated into a new population of entities. This 'artificial life cycle' is repeated until the properties of these entities are satisfactory.

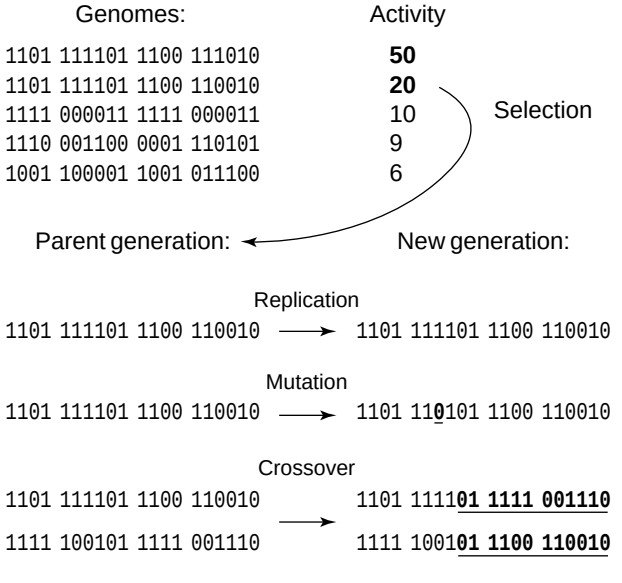


Figure 1. *Functions of a genetic algorithm: selection, replication, mutation and crossover.*

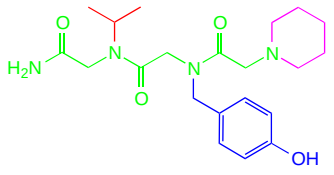
Optimization behaviour of GAs

Many parameters can be set and kept constant or varied during the course of a GA experiment: the size of the population, the number of surviving genomes, the crossover and mutation rate, the number of parents for the children and, finally, the ranking function. Finding optimal parameters for a given problem is an optimization problem in itself. The 'structure' of the search space also has a large influence on whether or not a GA will be successful[3,4]. The good search power of GAs is believed to originate from the building-block hypothesis[5,6], which states that the combination of 'fit' building blocks or contiguous schemes of genes on the genomes might yield higher-order schemes of even better fitness. This optimization behaviour perfectly matches the discontinuous, nonsteady structure space of chemistry, which is also formed by 'building blocks' (that is, atoms, reagents, starting materials and reactions); these building blocks are then statistically analyzed by the GA. Combinatorial compound libraries are generated even from systematic arrays of building blocks; one should thus expect to obtain a systematic SAR as well. The task of an optimization procedure is to predict, with a relatively low sampling, yet more active molecules.

Genes of small molecules

The first step in using GAs for small molecules is to invent a suitable encoding system for the chemical space of interest – the 'genes' of molecules. General structure-based algebraic representations were first developed by Ugi[7], who used matrices to describe the molecule as an array of atoms and shared electrons. More recently, Weiniger[8] introduced the SMILES notation (Fig. 2), which uses a string of letters to describe a molecule. Chemical rules that utilize our knowledge about how to connect atoms or groups of atoms (such as >C<, –H, –O– and >C=O) can be encoded in a similar way and used not only to describe molecules but also to build them[9,10]. These rules could include, for example, the chemical feasibility that either forbids the construction of unlikely four-atom combinations or limits the number of rotatable bonds.

Combinatorial libraries can allow an alternative, efficient encoding by enumerating only substituents or building



SMILES: NC(=O)CN(C(C)C)C(=O)CN(Cc1ccc(O)cc1)C(=O)CN2CCCCC2

Decimal: 1 2 1 88 1 101

Binary: 1001 1110 1111

Cyclic: –1001 1110 1111 01–

Figure 2. *Encoding of combinatorial libraries. The example shown is a tripeptoid-type structure, which allows various encoding schemes needed for genetic algorithms. Corresponding substructures and encoding schemes are marked by the same colour.*

blocks having a certain backbone structure. Arbitrary binary bit-strings[10,11] or decimal bit-strings[12] have been used to generate the genomes of molecules that have meaning only in the context of a particular compound library rather than to a general encoding of molecules with SMILES strings (Fig. 2). Using different reaction conditions, different backbone structures can be obtained with the same set of building blocks in one combinatorial library. In one particular case, the reaction products were encoded in a closed, circular binary genome (–1001 1110 1111 01–, where 1001, 1110 and 1111 encode the respective building blocks and 01 encodes the specific reaction conditions; see Fig. 2), as we described at the International Symposium on Combinatorial Chemistry held in Schliersee (Germany) in September 1996. These circular genomes have the further advantage that crossover takes place with the same probability for each bit, contrary to linear bit-strings for which the crossover probability is lower for the first and last bits of the string.

Diversity of compounds

The recent approach to measuring diversity is based on dissimilarity – the opposite of similarity. Similarity, in turn, is defined using problem-dependent criteria such as activity for a certain biological target. Thus, two nonactive molecules are dissimilar with respect to the active molecule, but they are not necessarily diverse. Moreover, two previously similar molecules might be found to be dissimilar in relation to another biological target. Thus, it is not possible to develop a general measure for biologically and structurally diverse molecules in order to generate a 'universally diverse' compound library for biological screening. However, within the context of either predefined selection functions or the predefined chemical space of combinatorial libraries, diversity can be defined and used to generate diverse or similar molecules within this limited measure. Thus, in nature, evolution generates diversity by changing an essential amino acid required for biological activity by a single mutation on the DNA; mutation at a less important site yields a similar protein. The statistics of this encoding provides a measure on the structure–activity of peptides and their biological function; it has been used by Schneider[13] to predict peptides with wanted biological activities using evolutionary algorithms and neuronal networks. Analogously, with respect to the selection function used, GAs could be used to generate diversity and similarity simultaneously in one population[5] by mutation and crossover.

Design of similar compounds

For general molecules, the combination of encoding molecules and statistical analysis has been used to determine common encoding schemes for dopamine D_2 ligands by statistical comparison[14]. A derived 'average' modal binary bit-string was used to search other databases for molecules having the same structural commonalties – this represents a new type of advanced substructure searching technique for similarity. New methods to predict biologically active similar molecules use neuronal networks in combination with GAs (Refs 15–17) to correlate biological activities with physicochemical parameters. In this case, it is not the molecules themselves but rather their parameters (including molecular weight, surface area, molecular dimensions, charges, polarity, lipophilicity, H-bond donor or acceptor properties, volume, pK_a values, polarizability, σ and π donors and acceptors, or flexibility) that are encoded, and a mathematical model using these parameters to describe the QSAR of interest represents the genome to be optimized. The generality and value of the models obtained, however, are limited to the biological activity selection function that was used to build the model.

GAs have been developed to select molecules from a large virtual library exhibiting structural similarities to a given molecular target molecule (e.g. a known drug). Tripeptoid-like molecules have been built[12] in the computer

by a GA choosing from a set of 2507 × 2507 × 3312 preselected building blocks giving a library size of ~20 billion. To test the concept and to study the GA's optimization behaviour in finding the optimum in this combinatorial problem, the specific tripeptoid shown in Fig. 2 was chosen as the target molecule from this library. A topological descriptor using atom pairs separated by a specific number of bonds was used as the selection criterion for similarity. Several GA parameter strategies were studied, such as the stochastic + best-third selection and random + neighbours mutation. In the stochastic selection procedure, parents are chosen randomly from the previous population to generate new children, whereas, in the best-third method, the top-scoring, best third of all parents is transferred unchanged, the worst third is eliminated and the medium third is used to generate new children via crossover. Random mutation permits each gene to be mutated with equal probability, whereas neighbours mutation follows a given rule that a mutation can lead only to a similar building block. With a population size of 300 molecules and the elitist best-third selection and neighbours mutation, the right answer was found in the above peptoid example after only 12 generations! This result is rather astonishing since only 3600 peptoids were examined out of the 20 billion. Known cholecystokinin and angiotensin converting enzyme antagonists were then chosen as molecular targets to search for similar tripeptoids. A striking structural similarity between the proposed peptoids and the target molecules was generally achieved after only 20 generations. Unfortunately, the biological activity of these proposed peptoids has not been reported.

A GA has been used in a similar way to propose new polymer molecules that mimic a given target polymer[9]. The molecules were built in the computer by linking predefined main-chain or side-chain atoms or atom groups, which were considered as the genes of the molecules, together with several chemical rules about stability. Some new interesting GA operators were introduced, such as insertion and deletion of genes into chromosomes, shifting main-chain atom groups into other positions on the chromosome, or blending parent chromosomes into one large chromosome. Even more chemical rules are needed when generating general, nonpolymeric molecules of all structural classes with a GA (Ref. 10). Target molecules with a given molecular weight and a three-dimensional shape were chosen as an example. The method was stated to be of use for any molecular target such as enzyme inhibitors, polymers or new materials.

Design of diverse compounds

An optimally diverse compound library for biological random screening would be a collection of representatives from a variety of compound clusters of similar molecules. Thus, the first step is to cluster molecules according to their similarity and then to compose a diverse set. The central task to solve this problem for all kinds of molecules is the identification of suitable molecular descriptors. A variety of these descriptors and clustering methods have been evaluated by Brown and Martin[18] for the design of optimal screening libraries. The simplest method, counting 153 different substructure MACCS keys in a molecule, provided a better similarity measure than using more-complex two- or three-dimensional criteria.

One may also choose a molecular property and select for molecules that are different with respect to that property. A dissimilar, unique molecular weight is such a property that would facilitate, for example, the deconvolution of combinatorial library mixtures by mass spectroscopy. The optimal design of such mixtures was the target function in a recent application of a GA (Ref. 11).

The diversity of combinatorial libraries appears to be more accessible by computational methods than general molecules are, because of their well-defined 'closed' chemical space. Thus, several methods have been introduced to design diverse combinatorial compound libraries by selecting optimal building blocks for synthesis[19,20]. Using molecular volume, lipophilicity, charge and H-bond donor or acceptor descriptors[21], it was shown that peptoid-type combinatorial libraries can be designed to exhibit the same density of structural fragments per atom[22] as commercial drugs. The implicit, but unproven, assumption is that one might thus also find all the desired biological activities within this library.

Evolutionary chemistry

The evolution of peptides by phage display libraries, combinatorial biochemistry or even artificial evolution of enzymes[23] has been established using the evolutionary mechanisms of nature together with a biological selection function such as a binding or enzyme assay. Evolutionary chemistry for small, nonoligomeric and general molecules has now become feasible by the application of GAs and encoding molecules from a combinatorial library in the computer. There are already examples reporting the successful connection of GAs, organic synthesis of proposed molecules and biological testing in an evolutionary chemistry feedback loop.

Peptidic enzyme ligands

A population of 24 randomly chosen hexapeptides was optimized to exhibit increased trypsin-inhibiting properties using a GA (Ref. 24). According to the best-third method described above, the best six peptides out of the 24 were duplicated after biological testing in a chromogenic assay with trypsin; the worst six were eliminated and the rest were kept constant to arrive at a new population of 24 peptides. This new population of peptides was then changed by a crossover rate of 100%, choosing two parents at random. Thereafter, mutation was applied with a probability of 3%, providing a GA with a slight elitism. The average inhibitory activity was improved from 16% of the first randomly chosen population to 50% in the sixth generation. Moreover, 13 of the 25 most active peptides comprised a consensus sequence of Ac-XXXXKI-NH_2 and eight out of these had a Ac-XXKIKI-NH_2 sequence. The best identified peptide was Ac-TTKIFT-NH_2, with an inhibition of 89%, being identical to a previously found trypsin inhibitor and demonstrating analogy to phage display library methods.

In another example, only 300 peptides were synthesized in five generations to obtain substrates for stromelysin out of a pool of 64,000,000 possible random hexapeptides[25]. The peptides were synthesized on a solid support with a fluorescence marker at the N-terminus. After proteolytic cleavage, the nonfluorescent beads were analysed. The starting sequence was biased towards using 60 peptides of the sequence $X_1PX_3X_4X_5X_6$, removing the bias in all subsequent generations. From a population of 60 peptides the best one was copied to the new generation and the others were changed by a crossover rate of 60%. The new peptides were then subjected to mutation, with a rate of 0.1% applied to each bit of the 30 bit gene, giving a 3% overall mutation rate. The GA was terminated when 95% of the population was identical. The hexapeptide MPQYLK was identified as the best substrate for stromelysin in the final generation, being cleaved between tyrosine and lysine. The selectivity of the new substrates compared with collagenase was also determined, and selective substrate sequences were identified for both enzymes. Thus, this method could help not only to find new substrates but also to obtain structure–activity and selectivity ideas for new inhibitors.

Evolutionary programming

Evolutionary programming (EP) is a technique that bears many similarities to GAs. In basic terms, unlike GAs, no crossover is used, and changes to the genomes are introduced in somewhat smaller steps (such as by series of mutations). Neuronal networks in combination with EP methods have been applied to the design of signal peptidase I (SP-I) cleavage sites[26,27]. This 'simulated molecular evolution' used a training set of 85 known protein sequences that are cleaved or not cleaved by SP-I to train an artificial neuronal network over 100 generations by using four selected amino acid properties: hydrophobicity, hydrophilicity, polarity and volume. The networks were then used to predict new leader peptidase substrate sequences.

Sequence-dependent SARs were discovered – for example, pronounced hydrophobic amino acids regions are needed for a sequence that can be cleaved by leader peptidase. Properties of the building blocks and their positions in the peptide sequence were thereby explicitly discovered by statistical analysis on the performance of the corresponding peptide – it is a major difference to pure GAs, which build up these relations implicitly. The peptide FFFFGWYGWARE was proposed as a consensus cleavage site, with a presumed cleavage between alanine and arginine. The peptide was then cloned in *Escherichia coli* into a fusion protein. The protein was indeed processed by the secretory proteases.

A similar recent example started with a 'seed' decapeptide that binds to the anti-1-adrenoceptor autoantibody[28]. A small set of 90 similar peptides was then generated, synthesized and screened for binding affinity. Using these results, a neuronal network was trained with an evolutionary algorithm to fit the observed results. New decapeptides were then predicted that had only one or two amino acids in common with the initial seed peptides. Subsequent synthesis and testing showed that both peptides were able to prevent the chronotropic effect caused by the autoantibodies, as expected.

Nonpeptidic inhibitors

We have already published the first example of nonpeptidic molecules for the development of thrombin inhibitors[29]. Using the well-known Ugi reaction and 10 × 40 × 10 × 40 building blocks, 160,000 combinations are possible (Fig. 3). Ten 'biased' amine building blocks at substituent R_3, which were known to bind with low affinity into the arginine P1 recognition pocket of thrombin, were preselected. Whereas in the initial population the best reaction product had an inhibitory activity of ~300 μM,

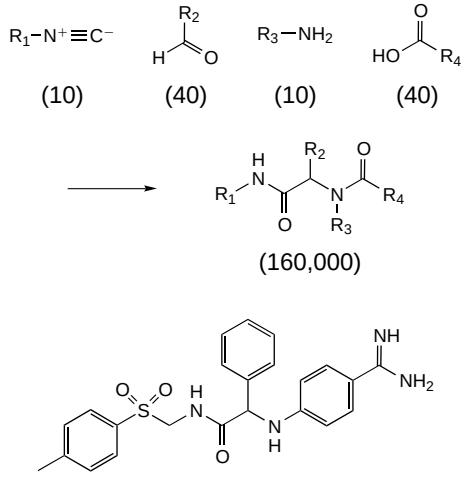


Figure 3. *The four-component Ugi reaction starting from 10 × 40 ×10 × 40 building blocks gives 160,000 possible combinatorial products. The* N*-aryl-phenylglycine amide derivative shown was found to be a submicromolar thrombin inhibitor from among the reaction products.*

a submicromolar thrombin inhibitor was discovered after 20 generations of 20 single Ugi products per population using a robotics system. Interestingly, this *N*-aryl-phenylglycine amide derivative (Fig. 3) was a three-component side product of the four-component reaction. In principle, the encoding was done for the corresponding educts (starting materials) of this reaction and not for the final expected products.

Contrary to all other reported GA procedures, all generated and tested reaction products were stored together with their biological activities in a database. Newly proposed compounds were then checked against this database to avoid the useless re-synthesis and re-testing of already known compounds. The 20 best compounds that were used to generate the new generation were selected from this database during the course of the experiment. Thus, the new generation is the result of all previous populations and not derived just from the results of the last generation. Whereas nature is able to 'remember' good performing molecules only via error-free replication and re-synthesis, the computational evolutionary approach allows the easy storage and use of all results of the GA experiment. The 20 bit-strings were then subjected to 100% crossover and 1% mutation to generate 20 new molecules after each biological test. First, the two top-scoring parents, then the two next best parents from this list were chosen for crossover and so forth, giving a strong basis on which to explore around an observed biological activity. This elitist optimization behaviour, however, is counterbalanced by the mutation that gives rise to large structural changes in these small molecules, preventing exhaustive exploration around suboptimal solutions.

Summary

All reported synthetic examples deal with the search for desired molecular properties in a chemical sequence space. This problem seems to be well suited to combinatorial optimization strategies using EP and GAs. Moreover, a properly set up GA should converge relatively quickly in 10–15 generations[30]; this was indeed observed in most of the pure computational examples and in the evolutionary chemistry examples. A series of different applications of GAs has been successfully introduced in chemistry, as well as in areas not mentioned in this review, such as molecular modelling[31] and optimization of reaction conditions for combinatorial chemistry[32]. However, for general problems, there is often little understanding of why a particular GA succeeded or failed. The relationship between the structure of a problem and the observed performance of GAs still requires research.

From the given examples it would seem that the structure of the search space is more important than fine-tuning the GA parameters. Combinatorial chemistry, which generates a systematic chemical space, seems ideally suited for novel computational methods such as GAs. The value and impact of these novel methods can be assessed almost instantly when connected with automated synthesis and screening. Thus, the idea of evolutionary chemistry yielding small drug-like molecules with desired properties in automated feedback cycles has been demonstrated. Whether this technique can be developed to the same maturity as phage display libraries or the generation of antibodies will become clear in the next few years, and it will depend on a closer integration of computation, synthesis and screening in the drug discovery process.

REFERENCES

1 Cook, W.J. *et al.* (1997) *Combinatorial Optimization*, Wiley
2 Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press
3 Rechenberg, I. (1973) *Evolutionsstrategie: Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*, Frommann-Holzboog

4 Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning,* Addison-Wesley
5 Holland, J.H. (1996) *Hidden Order – How Adaptation Builds Complexity*, Addison-Wesley
6 Forrest, S. and Mitchell, M. (1993) in *Foundations of Genetic Algorithms 2* (Whitley, D., ed.), pp. 109–126, Morgan Kaufmann
7 Ugi, I. *et al.* (1990) in *Concepts and Applications of Molecular Similarity* (Johnson, M.A. and Maggiora, G.M., eds), pp. 239–288, Wiley
8 Weiniger, D. (1988) *J. Chem. Inf. Comput. Sci.* 28, 31–36
9 Venkatasubramanian, V., Chan, K. and Caruthers, J. (1995) *J. Chem. Inf. Comput. Sci.* 35, 188–195
10 Glen, R.C. and Payne, A.W.R. (1995) *J. Comput.-Aided Mol. Design* 9, 181–202
11 Brown, R.D. and Martin, Y.C. (1997) *J. Med. Chem.* 40, 2304–2313
12 Sheridan, R.P. and Kearsley, S.K. (1995) *J. Chem. Inf. Comput. Sci.* 35, 310–320
13 Schneider, G. *et al.* (1995) *Minimally Invasive Med.* 6, 106–115
14 Shemetulskis, N.E. *et al.* (1996) *J. Chem. Inf. Comput. Sci.* 36, 862–871
15 So, S-S. and Karplus, M. (1996) *J. Med. Chem.* 39, 1521–1530
16 So, S-S. and Karplus, M. (1997) *J. Med. Chem.* 40, 4347–4359
17 So, S-S. and Karplus, M. (1997) *J. Med. Chem.* 40, 4360–4371
18 Brown, R.D. and Martin, Y.C. (1996) *J. Chem. Inf. Comput. Sci.* 36, 572–584
19 Good, A.C. and Lewis, R.A. (1997) *J. Med. Chem.* 40, 3926–2936
20 Gillet, V.J., Willett, P. and Bradshaw, J. (1997) *J. Chem. Inf. Comput. Sci.* 37, 731–740
21 Shemetulskis, N.A. *et al.* (1995) *J. Comput.-Aided Mol. Design* 9, 407–416
22 Martin, E.J. *et al.* (1995) *J. Med. Chem.* 38, 1431–1436
23 Reetz, M.T. *et al.* (1997) *Angew. Chem.* 109, 2961–2963
24 Yokobayashi, Y. *et al.* (1996) *J. Chem. Soc., Perkin Trans. I* 20, 2435–2437
25 Singh, J. *et al.* (1996) *J. Am. Chem. Soc.* 118, 1669–1676
26 Schneider, G. *et al.* (1995) *Minimally Invasive Med.* 6, 72–77
27 Wrede, P. *et al.* (1998) *Biochemistry* 37, 3588–3593
28 Schneider, G. *et al.* (1998) *Proteins* 30, 49–60
29 Weber, L. *et al.* (1995) *Angew. Chem., Int. Ed. Engl.* 107, 2453–2454
30 Wagener, M. and Gasteiger, J. (1994) *Angew. Chem., Int. Ed. Engl.* 33, 1189–1192
31 Devillers, J. (1996) *Genetic Algorithms in Molecular Modeling*, Academic Press
32 Almstätter, M. (1997) in *Proceedings of the German–Polish Workshop on Multicomponent Reactions in Combinatorial Chemistry*, 28–30 September, Rzeszow, pp. 167–180