Application of Genetic Algorithms to Combinatorial Synthesis: A Computational Approach to Lead Identification and Lead Optimization[†,▽]

Jasbir Singh,*[,‡] Mark A. Ator,[§] Edward P. Jaeger,[||] Martin P. Allen,[‡] David A. Whipple,[‡] James E. Soloweij,[§] Swapan Chowdhary,[⊥] and Adi M. Treasurywala[||]

Contribution from the Departments of Medicinal Chemistry, Biophysical and Computational Chemistry, Enzyme and Receptor Biochemistry, and Analytical Sciences, Sanofi Winthrop, 1250 South Collegeville Road, Collegeville, Pennsylvania 19426-0900

Received September 15, 1995[⊗]

Abstract: A genetic algorithms (GA) based strategy is described for the identification or optimization of active leads. This approach does not require the synthesis and evaluation of huge libraries. Instead it involves iterative generations of smaller sample sets, which are assayed, and the "experimentally" determined biological response is used as an input for GA to rapidly find better leads. The GA described here has been applied to the identification of potent and selective stromelysin substrates from a combinatorial-based population of 20^6 or 64 000 000 possible hexapeptides. Using GA, we have synthesized less then 300 unique immobilized peptides in a total of five generations to achieve this end. The results show that each successive generation provided better and unique substrates. An additional strategy of utilizing the knowledge gained in each generation in a spin-off SAR activity is described here. Sequences from the first generations were evaluated for stromelysin and collagenase activity to identify stromelysin-selective substrates. GlyProSerThr-TyrThr with Tyr as the P_1' residue is such an example. A number of peptides replacing Tyr with unusual monomers were synthesized and evaluated as stromelysin substrates. This led to the identification of Ser(OBn) as the best and most selective P_1' residue for stromelysin.

Introduction

Recently, combinatorial/multiple synthesis of both oligomeric and non-oligomeric libraries of diverse compounds and high-throughput screening have provided a format for the identification of new lead compounds for various molecular targets.[1] However, in any given template, the number of possible compounds one can synthesize in combinatorial or permutational[2] libraries is enormous, often in the millions.[1] Typically, one prepares libraries containing 10^4-10^6 compounds per template, assays these in a number of diverse screens, and subsequently identifies actives—"hits". High-throughput screening for a large number of targets is essential for success with this format. There have been a few attempts to evaluate the molecular diversity contained in a given library prior to synthesis.[3] The motivation of this diversity assessment approach is to select a subpopulation which maximizes dissimilarity[4] among the selected members. The synthesis and biological evaluation of this subpopulation rather then the entire library is a more manageable task. One of the important goals in combinatorial/multiple synthesis is to build high-fidelity libraries, which ensures greater probability of obtaining hits for a *given* biological target.[5] An approach to accomplish this objective would be to synthesize diverse compounds in smaller sets (say, in the low hundreds at a time) and utilize the biological response to guide the selection of compounds[6] for successive synthesis and biological evaluation (Scheme 1). Recently, genetic algorithms have been successfully utilized to find solutions to a number of complex problems.[7−9] In fact, genetic algorithms (GA) are distinguished for their powerful optimization characteristics, enabling them to find a set of very good (but not necessarily the best) solutions rapidly where an

[†] The work described here was carried out at Sterling Winthrop Pharmaceutical Research Division, before its divestiture on October 3, 1994, by Eastman Kodak to Sanofi Winthrop. Current Addresses: A.M.T., Allelix Biopharmaceuticals, 6850 Goreway Drive, Mississauga, Ontario L4V 1V7, Canada. E.P.J., 3 Dimension Pharmaceuticals Inc., 665 Stockton Drive, Exton, PA 19341. M.P.A. and D.A.W., Pfizer Central Research, Eastern Point Road, Groton, CT 06340. J.E.S., Amgen Inc., 1840 DeHavilland Dr., Thousand Oaks, CA 91320. M.A.A., Cephalon, Inc., 145 Brandywine Parkway, West Chester, PA 19380. S.C., Sanofi Winthrop, 31 Great Valley Parkway, Malvern, PA 19355.

[‡] Department of Medicinal Chemistry.

[§] Department of Enzyme and Receptor Biochemistry.

[||] Department of Biophysical and Computational Chemistry.

[⊥] Department of Analytical Sciences.

* Author to whom correspondence should be addressed at NYCOMED Inc., 466 Devon Park Drive, P.O. Box 6630, Wayne, PA 19087-8630.

[▽] Abbreviations: FMOC, (9-fluorenylmethoxy)carbonyl, CPG, controlled pore glass, AMP, ((aminopropyl)silyl)oxy, β-Ala, β-alanine, Acp, 6-aminocaproic acid, COP, 7-hydroxycoumarin-4-propionic acid, HOBt, *N*-hydroxybenzotriazole, mCl-t, a recombinant form of the human fibroblast collagenase, mSl-t, a recombinant form of the human fibroblast stromelysin. Single and three letter codes for 20 amino acids used for this work are as follows: A (Ala), D (Asp), E (Glu), F (Phe), G (Gly), H (His), I (Ile), K (Lys), L (Leu), M (Met), N (Asn), P (Pro), Q (Gln), R (Arg), S (Ser), T (Thr), U (denotes *S*-methylcysteine, Smc), V (Val), W (Trp), Y (Tyr).

[⊗] Abstract published in *Advance ACS Abstracts,* February 1, 1996.

(1) For an excellent review see: (a) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. *J. Med. Chem.* **1994**, *37*, 1233−1251. (b) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. *J. Med. Chem.* **1994**, *37*, 1385−1401 and references cited therein.

(2) Pirrung, M. C. *Chemtracts: Org. Chem.* **1994**, *7*, 184−186.

(3) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moose, W. H. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(4) See ref 3 for indices used to represent dissimilarity.

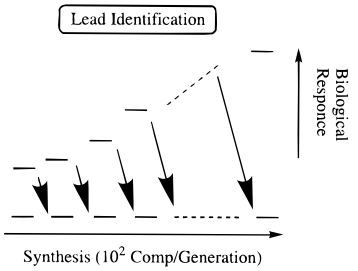
(5) There is tremendous effort in producing and screening chemically diverse compound libraries. However, the real interest for the pharmaceutical industry is the biological diversity that is embodied in it, i.e. *bio*diversity not necessarily *chemo*diversity.

(6) This approach would not require explicit description of a set of indices, but instead, the selection of indices would be implicit in this type of optimization approach.

(7) For general references for genetic algorithms see: (a) Holland, J. H. *Sci. Am.* **1992**, 66. (b) Forrest, S. *Science* **1993**, *261*, 872−878.

0002-7863/96/1518-1669$12.00/0

Scheme 1. Biology-Guided Lead Identification Paradigm[a]



[a] Arrows represent input of biological data for generation $Gen_{(i)}$ to guide selection of compounds for generation $Gen_{(i+1)}$ (see text for details).

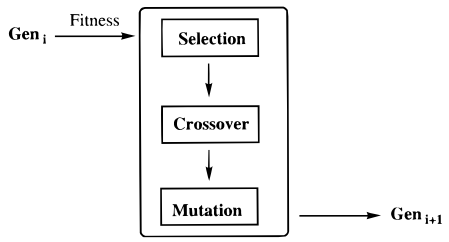
astronomically larger number of potential possibilities exists. Even though there have been numerous[10] applications of GA, to the best of our knowledge, there is no example of an application of GA to guide chemical synthesis for structure optimization for any class of compounds. In this paper, we report the first application of GA-guided chemical synthesis. Our principle criteria for the choice of a test case to explore the usefulness of GA to guide chemical synthesis was to choose a template for which chemical synthesis has been well established, so that we could evaluate GA's impact to find "hits" without confounding it with synthesis-related issues. Therefore, a peptide-based template was our first logical choice. Recently, we have reported the screening of immobilized peptide libraries as a tool for the determination of substrate specificity and selectivity for proteases.[11] We reasoned that selection (and optimization) of hexapeptides consisting of 20 amino acids,[12] representing 20^6 (64 000 000) possible structures, possessed all of the essential ingredients to be a good initial area in which to test the concept. The number of possibilities in the entire library was very large. The synthesis of these peptides had been previously worked out.[11] A validated assay was in hand and available.[11]

Methods

Genetic Algorithms. In this section, we will describe the basic ideas of the GA method, some issues involved in its use as a tool for the selection and representation of chemical structures, and finally the details of our implementation.

GA optimization methods are based on several strategies from Darwinian theories of evolution. In the normal survival and evolution of the species, new genetic mutants constantly arise and their survival and "dominance" is based on their ability to find food, reproduce, and resist "assault" on their existence. These would be classified in the language of genetic algorithms as the "objective" function which is being optimized. In the same way, if our living mutating population was made up of hexapeptides instead of species of organisms and the evolutionary pressure being applied (the biological function) was the biological activity, then one could envisage exactly the same process occurring. The GA we use here is based on three basic strategies:[13] *selection*, *crossover*, *and mutation*. The first of these strategies, *selection*, is the use of a breeding population in which the individuals who are more "fit" in some sense (higher biological response in this application) have a higher chance of producing offspring and passing on their "genetic" information. The second strategy is the use of *crossover*[14] (mating) in which a child's genetic material is a mixture of his or her parents'. The final strategy is that of *mutation*, where the genetic material is occasionally "corrupted" to maintain a certain level of spontaneous and random genetic mutation in the population.

Scheme 2. Summary of Variables Used for Genetic Algorithms



POPULATION SIZE (N_{pop}) = 60

GENOME SIZE = 30 bit string {[6 monomers] x [5 bits (N_b) per monomer]}

NUMBER OF PARENTS COPIED UNCHANGED to Gen $_{(i+1)}$ = 1

(parent with the highest fitness is the only one copied as such)

CROSSOVER RATE = 0.6 (i.e. 60% of the selected members are mated in pairs)

MUTATION RATE = 1 in 1000 bit flip (from '0' to '1' or '1' to '0')

(frequency of randomly inverting a bit somewhere in the genome)

FITNESS VALUE = - fluorescence value (see text for details)

TERMINATION CRITERIA = 0.95

(i.e. 95% of the bits in a given generation are identical)

The GA paradigm used here employed a modified version of the Genesis GA[15] code and is outlined above in Scheme 2. We work with a population of individuals which interact through their genetic operators to carry out an optimization process. An individual is specified by a chromosome, a bit string in this case. Let us assume that a hexapeptide is to be represented by a bit string (i.e., a sequence of 1's and 0's) of 30 bits (or digits). Each amino acid is then represented by five bits: the first amino acid being coded into bits 1−5, the second being coded into 2−10, etc. Each five-bit code can essentially code for 2^5 or 30 unique amino acids. Since there are only 20 amino acids, this five-bit codon can easily accommodate a unique pattern for each amino acid.[16] Therefore, the 30-bit string can be translated into a unique hexapeptide (and vice versa). A fitness function, also called the objective function (see above), is used to rank the individual's chromosome. The optimization proceeds because the population produces individuals that have increasingly higher fitness. Initially, a set of N_{pop} individuals is formed by choosing a set of N_b-bit strings at random and each member is synthesized and evaluated for fitness. A roulette wheel is conceptually created where the "slice" on the wheel for any given individual is proportional to the value, for that individual, for its fitness. Biologically more active peptides in our implementation get a large slice in the wheel and inactive peptides get a small slice. In the selection process one may imagine mating pairs to be selected by spinning this wheel. (Note: *ALL* individuals have a place on the wheel and therefore have a finite chance to be selected). This produces a list of pairs for mating.

Subsequent generations are formed as follows: each member of the first generation[17] is ranked by fitness, and the fittest individual is placed into the next generation with no change. Next, pairs of individuals (from the selection step above) are crossed-over to form the next generation. The crossover step may be visualized as follows (although

(8) (a) Wagener, M.; Gasteiger, J. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 1189−92. (b) Walters, D. E.; Hinds, R. M. *J. Med. Chem.* **1994**, *37*, 2527−2536. (c) Wehrens, R.; Lucasius, C.; Buyden, L.; Kateman, G. *Anal. Chim. Acta* **1993**, *277*, 313−324. (d) See ref 9 in ref 6a (listed above) for use of genetic algorithms for jet engine design.

(9) (a) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. *J. Comput. Chem.* **1993**, *14*, 1407−1414. (b) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M. *J. Mol. Struct. (THEOCHEM)* **1994**, *308*, 191−206.

(10) See refs 10−16 in ref 8a above.

(11) Singh, J.; Allen, M. A.; Ator, M. A.; Gainor, J. A.; Whipple, D. A.; Soloweij, J. E.; Treasurywala, A. M.; Morgan, B. A.; Gordon, T. D.; Upson, D. A. *J. Med. Chem.* **1995**, *38*, 217−219.

(12) We use all 20 coded amino acids (see a complete list under Abbreviations), except Cys. We employ *S*-methylcysteine (Smc, denoted by the single letter code U) as the 20th amino acid.

(13) We do not use deletion and insertion as we do not want to change the overall size of the chromosome and, therefore, overall length of the bit strings.

(14) Crossover is the single most important aspect which provides for most optimum assurance to explore the gene population for selecting a set of more fit members.

(15) In a true sense this random generation should be referred to as generation 0 (zero) as far as GA's are concerned, since there are no fitness functions which need to be evaluated by GA to provide the initial population of members.

(16) Genesis version 1.2 from ftp site: ftp.aic.nrl.navy.mil.

some of the actual details of the implementation are slightly different for technical reasons). The genome for each individual is of fixed length (in our case 30 bits). One can envision linking up the genomes for the two members of a mating pairs and then ARBITRARILY making a "cut" at a randomly chosen spot in both of the genomes. Recombining the first part of the first genome with the second part of the second and vice versa generates two "new" offspring individuals. This entire process is called crossover step. It is important that the total number of individuals N_{pop} selected for the subsequent generations remain identical to the initial random population, since each pair of parents produce exactly two offspring. After applying the selection and crossover steps as outlined above and thus producing a population of "new" individuals for the next generation, the mutation operator is applied. This simply consists in our case of "flipping" a bit (from 0 to 1 or vice versa). The frequency of this mutation is preset and constant throughout the run. The choice of which individual to mutate and which bit in that individuals' genome to mutate is purely random.

Chemistry. As stated earlier, we have selected a problem of protease substrate specificity and selectivity determination as a test case to evaluate the suitability and usefulness of GA in this area. A hexapeptide could be represented by the generic formula[18] $X_1X_2X_3X_4X_5X_6$. Early input into the makeup of the starting population for these study was based on the previously known result that proline in position 2 of a hexapeptide increased its chance of being a substrate of the target enzyme stromelysin.[11] It is normal when designing an assault on any biological target to use as much information as is available. Thus it was felt that fixing X_2 = proline for the selection of the initial "random" population set of 60 hexapeptides was a reasonable approach to biasing toward an early convergence. If this choice was not made, we reasoned that the initial population would probably all be inactive and no reasonable selection criteria could then be applied. It is however important to point out that this constraint was applied ONLY to the initial choice. Subsequent generations were free to chose non-proline amino acids for position 2. The initial random population was selected from a possible of 20^5 (3 200 000) possible hexapeptides (represented by $X_1PX_3X_4X_5X_6$). However, the X_2 = Pro constraint was not imposed on subsequent generations. The peptides were synthesized using controlled-pore glass as a solid support as described previously.[19] Controlled-pore glass (CPG) containing an ((aminopropyl)silyl)oxy (AMP) handle was exhaustively coupled with FMOC-β-alanine (β-Ala) followed by couplings with FMOC-ϵ-aminocaproic acid (Acp) using a 10-fold excess of preformed HOBt active esters in *N*-methylpyrrolidone. A 200 g sample of CPG containing a homogeneous population of linker [$(Acp)_5$-βAla] was prepared. The homogeneity of FMOC-$(Acp)_5$-βAla-CPG bulk sample was verified at each step of the coupling reaction by triplicate amino acid analysis using βAla as an internal standard. This sample was subsequently used for the preparation of the required peptides for any given generation. A standard protocol of triple coupling with a 10-fold excess of FMOC amino acid HOBt active esters (in situ activation method) was utilized for automated synthesis employing an Advanced ChemTech synthesizer model MPS 350. Finally, each peptide was reacted with FMOC alanine[20a] followed by capping with coumarinpropanoic acid (COP) as a fluorescent tag.[20b] A glass-bound peptide sample[21] used for biological assays could be generically represented as COP-A-$X_1X_2X_3X_4X_5X_6$-$(Acp)_5$-βAla-AMP-CPG, where X_i (i = 1−6) represents one of the 20 possible amino acids (see Abbreviations for the list of amino acids used).

Biological Assays. Automated biological assays were performed on a small sample of the glass-bound peptides in a 96-well format. Typically, 4.0 ± 0.3 mg of 62 glass-bound peptides were weighed[22] in individual tubes using the HP ORCA robot and 220 μL of a buffer (containing 50 mM Tris (pH 7.5), 200 mM NaCl, and 10 mM $CaCl_2$) followed by 44 μL of the protease solution (20 nM mSl-t or 10 nM mCl-t)[23] was added using a Packard PROBE. Samples were mixed on a variable speed vortexer for 120 min. The substrate was allowed to settle by gravity, a 125 μL aliquot of each sample was transferred to the appropriate position of a 96-well plate, and fluorescence was read using a fluorescence plate reader.

The stromelysin construct utilized has some autocatalytic activity. The positive control sample (GPLAMF) and a negative control sample (hexa-D-alanyl) were synthesized and evaluated as part of each generation. The assay results for these samples were used as an indication of the validity of the assay and to account for variance in stromelysin activity due to autocatalysis. The negative control samples typically produced very low fluorescence (<200) in the supernatant. The observed (raw) fluorescence for the positive control samples for generations Gen-1 to Gen-5 were 8652, 2959, 12 095, 15 242, and 13 044, respectively. The value for Gen-2 indicated the effect of the autocatalytic activity of this enzyme construct. In order to minimize the impact of the enzyme autocatalytic activity on the assays, we prepared a fresh sample of the stromelysin for each assay, by appropriate dilution of a bulk stock solution, just prior to assays. The fluorescence values for positive control samples for Gen-3 to Gen-5 screening supports this. We have recently reported an excellent correlation (r^2 = 0.994) between the relative k_{cat}/K_m ratio for soluble peptides and the relative substrate activity of corresponding immobilized peptides.[11] This correlation demonstrates that the kinetics of hydrolysis of immobilized peptides are predictive of the reaction of their soluble counterparts, validating the use of immobilized peptides for high-throughput screening.

Identification of Active Samples. Once the active samples (fluorescence value greater than control) were selected, the identity of these samples was validated by duplicate amino acid analysis. Next, a portion of the *post-enzymology* glass-bound sample was subjected to Edman microsequencing to identify the site of processing. A sample of the corresponding supernatant was used to obtain amino acid and mass spectral[24] analyses to identify the structure of the soluble fragment. The latter confirms the identity of the sample and the site(s) of processing by the protease. The assay format and approach for the identification of active sequence(s) are summarized in Scheme 3.

(17) Since we are using binary bit strings (0's and 1's), in order to represent 20 amino acids, we need 2^5 (i.e., 32) bits. Some of the 20 amino acids are represented by more then one string. The choice of this bit degeneracy was selected at random, but once selected, it was kept constant through out the experiment.

(18) Where X represents one amino acid present at a time and the numbering is used for the sake of discussion only.

(19) (a) Ator, M.; Beigel, S.; Dankanich, T.; Echols, M.; Gainor, J.; Gilliam, C.; Gordon, T.; Koch, D.; Kruse, L.; Morgan, B.; Olsen, R.; Siahaan, T.; Singh, J.; Whipple, D. *Peptides: Chemistry Structure and Biology*; Proceedings of the 13th American Peptide Symposium; Hodges, R., Smith, J., Eds.; 1994; pp 1012−1016. (b) We have previously described the reasons and the validation for use of CPG as a solid support, see ref 10 above for details.

(20) (a) We have incorporated an alanine residue at the N-terminus of all sequences identified by GA before we tag the N-terminus with the fluorescent marker—COP. This was carried out to distance the marker group further away from the active site of a protease. (b) Gainor, J. A.; Gordon, T. D.; Morgan, B. A. *Peptides: Chemistry Structure and Biology*; Proceedings of the 13th American Peptide Symposium; Hodges, R., Smith, J., Eds.; 1994; pp 989−991.

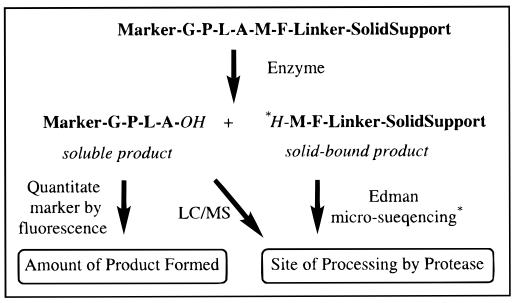
(21) The synthesis outlined here does not involve any "mix & split" strategy, and therefore, each sample/tube represents a single sequence.

(22) The weighing robot was programmed such that any sample which falls outside this weight range was reweighed. Each rack of assay also contained a negative control, hexa-D-alanyl, and GPLAMF as a positive control sample. These samples are used as a guide to decide the validity of every assay and are used as an integral part of biological evaluation.

(23) (a) Chowdhury, S. K.; Vavra, K. J.; Brake, P. G.; Banks, T.; Falvo, J.; Wahl, R.; Eshraghi, J.; Gonyea, G.; Chait, B. T.; Vestal, C. H. *Rapid Commun. Mass Spectrum.* **1995**, *9* (7), 563−569. (b) Brownell, J.; Earley, W.; Kunec, E.; Morgan, B. A.; Olyslager, B.; Wahl, R. C.; Houck, D. R. *Arch. Biochem. Biophys.* **1994**, *314*, 120−125.

(24) (a) Eshraghi, J.; Chowdhury, S. K. *Anal. Chem.* **1993**, *65*, 3528−3533. (b) Following is a typical sample procedure for obtaining data from LC/MS: The procedure adopted for on-line separation using a microcapillary HPLC system coupled to an electrospray ionization (ESI) mass spectrometer has been described in details in the above reference. Briefly, the gradient mobile phases (0.1% aqueous trifluoroacetic acid (TFA) and 0.1% TFA in acetonitrile) from the Waters 600 HPLC pump (200 μL/min) is split with a ratio of 100:1. The smaller fraction (2 μL/min) passes through a 0.5 μL injection loop followed by a microcapillary column (VYDAC C-18, 300 Å; 300 mm id × 15 cm) to first to the microdetection cell of a Spectroflow UV detector and then to the electrospray ionization chamber of a Finnigan TSQ 700 mass spectrometer (Finnigan Mat, San Jose, CA), while the larger fraction goes to the waste. The UV detector and the ESI mass spectrometer are operated in series so that measurement of the UV chromatogram and the mass spectra can be performed on the same effluents. A sheath liquid (2-methoxyethanol) was added to the LC effluents prior to electrospray ionization at a flow rate of 2 μL/min to assist the electrospray ionization.

Scheme 3. Protease Assay Format and Strategy for Identification of Actives



Results and Discussion

Identification of Active Samples—"Hits". Stromelysin (mSl-t) was used as the protease of choice. The recombinant version of mSl-t utilized in these experiments has some autocatalytic activity. This could further complicate the raw biological data (the observed fluorescence) as a number of assays were performed over a period of time. We have used a set of controls as references to validate a given screening run. The peptide GPLAMF is employed as the positive reference peptide, and all the data shown are normalized versus this reference sample.[25]

The negative value[26] of the observed fluorescence of given samples[27] of generation $Gen_{(i)}$ is used as input for optimization in the GA to provide sequences for the subsequent generation $Gen_{(i+1)}$. On the basis of the paradigm described in Scheme 1, one way to evaluate the usefulness of GA for lead selection and lead optimizations would be to demonstrate that (i) samples with greater activity are observed in the later generations and (ii) new "actives" are identified in later generations which are not present in earlier generations. The latter would provide an indication of the ability and effectiveness of the GA technique to explore diversity space. A plot of the average fluorescence[28] (activity) versus generation is shown in Figure 1. [The small number of generations reported in this paper are the result of the untimely closure of the research facilities where this work was being carried out. However, the authors believe that the method was so compellingly efficient and the findings so important to the chemical community at large that they have decided to make this report.] One can clearly see that the trend to greater activity, even in this relatively small number of generations evaluated, is evident. Not only did the average activity per generation improve (as highlighted in Figure 1) but each generation also identified *new* sequences with greater activity compared to the previous generations (see Figure 2). The graphical representation of the data in Figure 2 provides evidence for GA's capability to explore diversity space. It should be clear that the greatest impact on diversity exploration is due to "crossover" terms in these algorithms, since the mutation rate of 1 in 1000 bits is not an effective mode to increase diversity. Examination of sequences for the five generations also showed that, even though we constrained X_2 to proline *only* for the initial generation of "random" sequences, we did not see any variance at the X_2 position throughout these limited number of generations. In the future, a better approach to provide such constraints would be to use a higher bias for a given amino acid rather then an absolute constraint as utilized for the current experiment. This approach would be analogous to the protocol used in phage display based experiments.[29]

Multiple assays per sample at the early screening stage are not relevant from a high-throughput screening point of view. It should be emphasized that we have carried out a single assay per sample and the observed (raw) fluorescence has been used

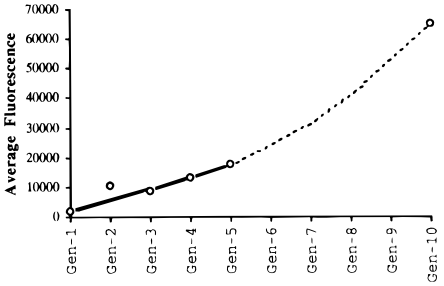


Figure 1. Performance evaluation of GA-based lead identification plot of average activity (fluorescence) versus GA-based generations. Solid lines represent actual experimental data, the dotted line shows a projected plot if one assumes that Gen-10 represents a termination point, each sample of Gen-10 being a good substrate with ~20% substrate processing.

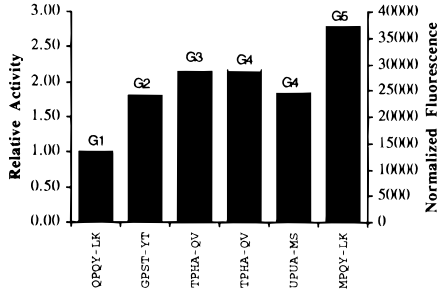


Figure 2. Most active sequence from each generation. The site of processing by stromelysin is indicated by a hyphen for sequences listed. Relative activity is plotted with best sequence for G1 = 1.0. Two samples are shown for G4. Even though the most active samples for G4 and G3 have identical sequences, the next most active sample for G4 represents a different sequence and very distinct P_2 and P_1' residues than observed for G3.[30]

(25) The observed (raw) fluorescence values were used as a fitness function. The normalized data have been used here to compare various generations. The data were normalized such that the activity of each positive control sample represents 10 000 fluorescence units.

(26) The genetic algorithm we have used here is very similar to the one utilized previously for computational experiments related to conformational analysis problems, where the GA's goal was to identify conformations with lower energy (i.e., the fitness function was used to compute energy); therefore, for this experiment, we have used the converse, i.e. negative, value as a fitness function to optimize activity.

(27) However, in a given generation (except of course the initial "random" generation), there are invariably some sequences which are already present in earlier generation(s). We have decided to re-test these samples for the number of occurrences asked for by GA. These multiple results take into account the variance in the biological assays. In addition, these results also provide an indication of the robustness and reliability of our assay.

(28) Average fluorescence is simply the sum of the normalized fluorescence value for all samples divided by 60 (total number of samples N_{pop} in any generation).

(29) For example, a 60−70% probability for occurrence of a given group should provide a better method to introduce constrains. This is analogous to the approach used for phage display based experimental design for introducing constrains at a given position(s) by using nucleotides NNN, where N represents 70:10:10:10 proportions of different nucleotide monomers. For a representative reference on phage display approach to introduce constrains, see: Schatz, P. J. *Biotechnology* **1993**, *11*, 1138−143.

(30) It is interesting to note that the sequences UPUAMS and UPUANS (Table 2, lines 16 and 19, respectively) are processed very differently by stromelysin. The sequence containing methionine (M) was cleaved at a single site, between alanine (A) and methionine (M). This cleavage pattern is consistent with the other samples for which the cleavage site was determined such that in all of these sequences proline occupies the P_3 pocket of the enzyme (see legend in Figure 2 for information on representative samples). However, the sequence containing asparagine (N) was cleaved at two distinct sites: P−U and A−N in a ratio of 3:1, respectively, where the major site of cleavage places the proline residue at the P_1 site.

Table 1. Representative Fluorescence Values (± Variance)[a]

sequence	Gen-1	Gen-2	Gen-3	Gen-4	Gen-5
APUELG	12116[b]	19660 ± 390	–[c]	–	–
APAELG	–	23508	14642 ± 502	14273 ± 831	16409 ± 1177
GPSTYT	9547	24138 ± 627	13465 ± 1139	12823 ± 145	19359 ± 40
MPGLUS	–	15194	12933 ± 318	–	–
MPELUS	–	–	14133	13170 ± 745	–
MPQYUS	–	–	–	21844	28091 ± 309
QPQYLK	13659	15708 ± 477	18431 ± 3326	18769 ± 318	25434 ± 1046
TPHAQV	–	–	29237	23385 ± 1193	34712 ± 1106
UPUANS	–	–	18079	15543 ± 590	21240 ± 253

[a] The fluorescence values shown have not been corrected for the actual weight of the sample used in the assay. As described in the Biological Assays section, the weights of the samples were 4.0 ± 0.3 mg, which may account for up to ±10% of the variability in the fluorescence value. Results were normalized to the fluorescence value of the standard GPLAMF sequence, as described in the text. [b] The single value indicates the first occurrence of this sequence. [c] Dashes indicate absence of this sequence in the given generation.

as a fitness function to drive GA-based optimization. This assures that every sample receives identical biological evaluation. It should also be stressed that the $Gen_{(i+1)}$ sequences are directly derived from the biological data obtained for the preceding generation, $Gen_{(i)}$. The absolute activity for a given sequence from run to run may vary, but the data within a given generation are reproducible (see Table 1). For example, sample QPQYLK has some variance in activity between different generations, but within any given generation, the activity for this sample is reproducible.

It is also important to realize that as part of the normal GA process a given sequence is identified several times. We had made a fundamental decision at the start of the GA-driven experiment to conduct biological evaluations for that sample the equal number of times as asked for by the GA and use the individual biological responses as a fitness function for the subsequent cycle. This we believed a way to handle the inherent variance in any biological evaluation.

A summary of results for selected samples[31] from five generations of the genetic algorithms (GA) based hexapeptides is represented graphically in Figure 3. The data in Figure 3 are arranged such that the location of a bar represents the identity of a particular sample (a specific sequence) and the patterns in the bars represent different generations. For example, the panel for the first generation highlights three distinct active peptides. The panel for the second generation shows the three samples which were present in generation one and four new active samples. Data for generation five clearly show several *new* sequences (shown by solid bars) not identified in the previous four generations. In addition, this panel (for Gen-5) also shows existing sequences highlighted by their respective generation pattern codes. Sequences in the order of their occurrence in Figure 3, along with the number of their multiple occurrences in the particular generation, are shown in Table 2.

We have used a single, simple fitness function—enhancement of the fluorescence value.[32] One may choose to incorporate a number of interdependent parameters or molecular properties as a fitness function, or one may employ penalty functions for some variables as a part of the fitness function paradigm. For example, one may construct complex fitness functions incorporating the molecular weight of the hexapeptide, its selectivity for the target enzyme over the related enzymes, its degree of overall charge, solubility in water, etc. The only requirement of this overall fitness function is that a value for each member can be determined. Many optimization strategies are inapplicable to the SAR problem because they require a continuous, mathematically definable fitness function to which analytic strategies may be applied. This is not the case with GA, which readily optimizes discontinuous fitness functions by its very nature. The method also readily accommodates variability in the fitness value (as seen here by using the raw fluorescence data). This is an essential part of SAR work and is also not readily accommodated by other methods of optimization. In fact, it is stated that the GA's succeed best where the fitness functions are highly complex, discontinuous, and "noisy".[7a] GA could easily be used to optimize a number of parameters simultaneously as it is known to provide solution to the kind of problem not suited to other methods. The combinatorial chemistry based diversity assessment/solution is a such a problem.

Identification of Selective Samples—"Unique" Hits. Finally, we emphasize that the entire process of iteratively optimizing some SAR has an inherent potential advantage over that of making large libraries before screening them. That advantage is one of "data digestion". In other words, the project team has an opportunity to assess results from one generation of compounds and to "spin-off" other avenues of investigation while the optimization is proceeding by the GA technique. We have found this to offer a great advantage in our hands, and an example of such a "spin-off" bonus is presented below.

Utilization of the samples from generation one for determination of stromelysin-selective substrate sequences, described below, is an illustration of the flexibility of this approach. However, it is important to emphasize that this activity *must* proceed in parallel with the GA optimization and not be used to influence its convergence. Thus one *should not* change the variables for the GA experiment on the basis of the outcome of information from a given generation. The goal in the stromelysin project was not only to identify good substrates but also to identify stromelysin-selective substrates. The substrate selectivity information would be subsequently translated into a selective inhibitor. In view of this objective, we have also assayed the initial set (generation 1) vs collagenase (mCl-t). The results for the collagenase vs stromelysin assays are shown below in Figure 4. The site of processing, in the sequences highlighted, is indicated by a hyphen. However, translation of these subtrates to a known class of matrix metalloprotease inhibitors[33] would only involve the P_1' and P_2' portion of the information. It is known that the P_1' residue imparts a greater selectivity among matrix metalloproteases. We have identified a sequence, GPST-YT, which is selectively processed by stromelysin, as shown in Figure 5. This processing between Thr and Tyr represents a unique selectivity between these two metalloproteases. Therefore, we used this stromelysin-selective substrate[34] and prepared a focused set of sequences which explores variations at the Y (Tyr) position (Chart 1) in an independent study. The groups were selected to evaluate a

(31) For sake of clarity only selected samples are shown. A complete list of sequences and their normalized data is available as supporting information.

(32) Since we have determined the site of processing only for a handful of samples in any given GA-based generations, we have not used site of cleavages as an input for obtaining the $Gen_{(i+1)}$ set in this study.

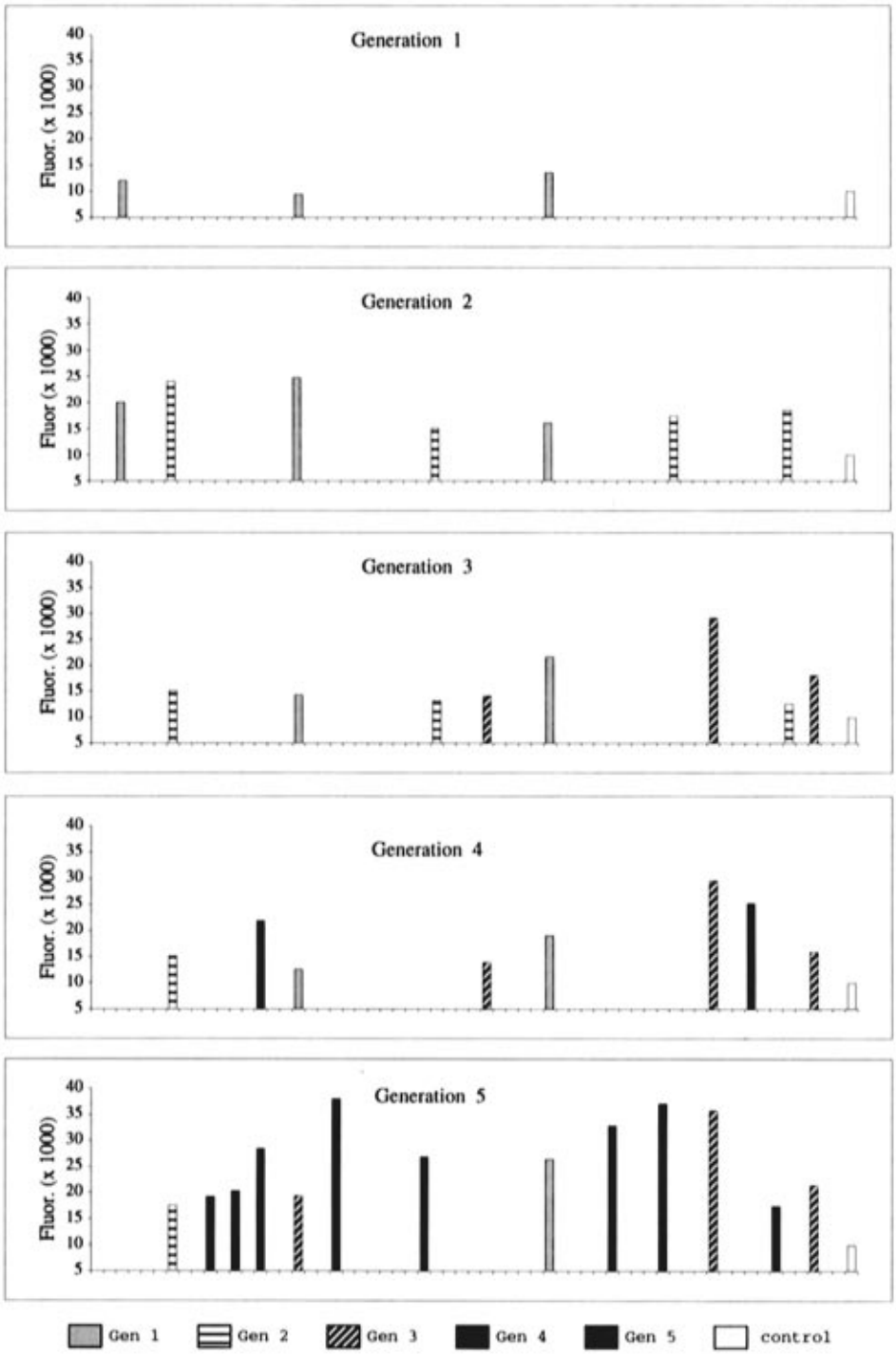


Figure 3. Data are arranged such that the location of a bar represents the identity of a particular sample (a specific sequence), and the pattern of the bar signifies the generation in which the sequence was discovered. The positive control sample, GPLAMF, used to normalize data for all generations is shown by the white bar in each panel. A listing of the sequences (in the order of their occurrence) in each generation (panel) is shown in Table 2.

variety of electronic and steric properties for their effect on relative hydrolysis of the substrates and thus their relative importance on overall binding energy in the active site of stromelysin. These immobilized samples were synthesized and assayed versus collagenase and stromelysin. The groups chosen are shown in Chart 1, and the assay results are shown in Figure 5. All of these samples demonstrate selective processing by stromelysin (mSl-t), as none of these showed any processing by collagenase (mCl-t). For three samples, shaded differently and marked with relative activity vs reference sample Tyr, we have confirmed by amino acid and LC/MS analyses[35] that the identity of the species in post-enzymology supernatant is COP-

(33) Various classes of MMP inhibitors have been reported in the literature. (a) For hydroxamate series, see: (i) Singh, J.; Conzentino, P.; Cundy, K.; Gainor, J.; Gordon, T.; Johnson, J.; Morgan, B.; Whipple, D.; Gilliam, C.; Schneider, E.; Wahl, R. *BioMed. Chem. Lett.* **1995**, *5*, 537−542. (ii) Johnson, W. H.; Roberts, N. A.; Borkakoti, N. J. *Enzyme Inhib.* **1987**, *2*, 1−22. (b) For *N*-carboxyalkyl series, see: Chapman, K. T.; Kopka, I. E.; Durette, P. L.; Esser, C. K.; Lanza, T. J.; Izquierdo-Martin, M.; Neidzwiecki, L.; Change, B.; Harrison, R. K.; Kuo, D. W.; Lin, T.; Stein, R. L. *J. Med. Chem.* **1993**, *36*, 4293−4301. (c) For phosphonate series, see: (i) Bartlett, P. A.; Marlowe, C. K. *Biochemistry* **1987**, *26*, 8553. (ii) Bird, J.; DeMallo, R. C.; Harper, G. P.; Hunter, D. J.; Karran, E. H.; Maekwell, R. E.; Miles-William, A. J.; Rahman, S. S.; Ward, R. W. *J. Med. Chem.* **1994**, *37*, 158−169.

(34) The sequence: GPST-YT, chosen for P1′ variations was observed in all five generations. This sequence happens to be processed selectively by stromelysin.

Table 2. Summary of Selected[a] Data Highlighting Unique[b] and Existing Sequences for Each Generation, as Identified by GA

no.	Gen-1	Gen-2	Gen-3	Gen-4	Gen-5
1	**APUELG**	APUELG (5)[c]	–	–	–
2	–	**APAELG**	APAELG (5)	APAELG (6)	APAELG (4)
3	–	–	–	–	**APQYUS**
4	–	–	–	–	**APYNLG**
5	–	–	–	**MPQYUS**	MPQYUS (4)
6	**GPSTYT**	GPSTYT (4)	GPSTYT (7)	GPSTYT (3)	GPSTYT (2)
7	–	–	–	–	**MPQYLK**
8	–	–	–	–	**QPHAQV**
9	–	**MPGLUS**	MPGLUS (3)	–	–
10	–	–	**MPELUS**	MPELUS (5)	–
11	**QPQYLK**	QPQYLK (7)	QPQYLK (4)	QPQYLK (6)	QPQYLK (2)
12	–	–	–	–	**TPHAQS**
13	–	–	–	–	**UPUAMT**
14	–	**SPYMEA**	–	–	–
15	–	–	**TPHAQV**	TPHAQV (7)	TPHAQV (6)
16	–	–	–	**UPUAMS**	–
17	–	–	–	–	**UPUAYT**
18	–	**TPLKSV**	TPLKSV (2)	–	–
19	–	–	**UPUANS**	UPUANS (5)	UPUANS (2)

[a] Sequences shown above are in the order of their occurrence in Figure 4 for each generation. Dashes indicate absence of that particular sequence for the given generation. [b] Unique sequences in each generation are shown in bold. [c] Number of times this sequence repeated in this generation.

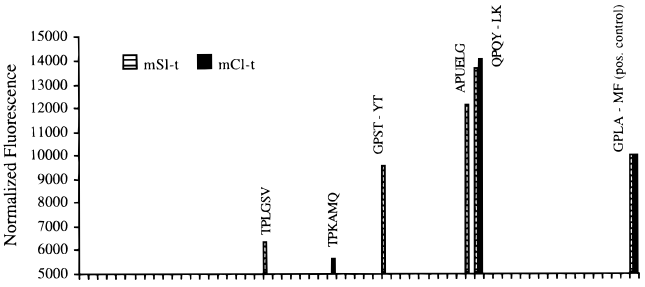


Figure 4. Assay results for collagenase (mCl-t) and stromelysin (mSl-t) for Gen-1 samples.

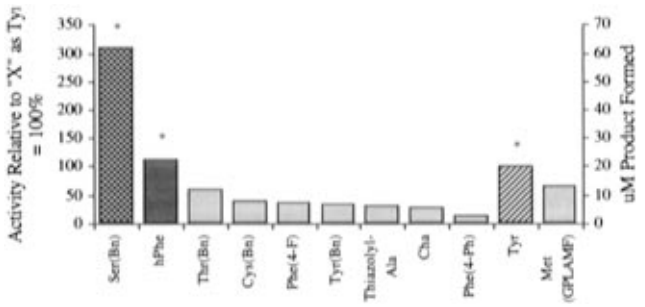
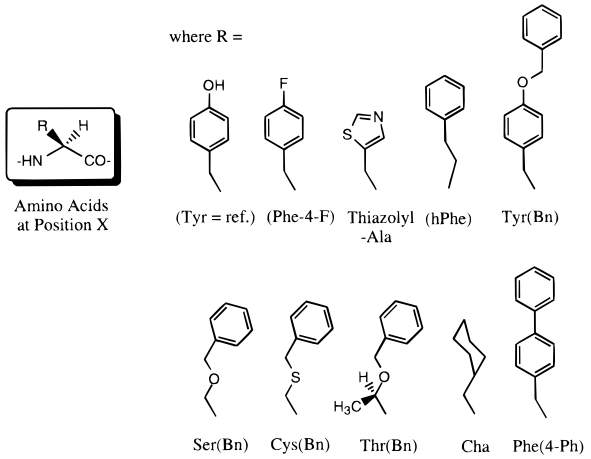


Figure 5. Assay results of samples COP-AGPST "X" T-(Acp)5-bAla-AMP-CPG. All of these samples showed processing by stromelysin only, and the data are shown above. Asterisks (∗) indicate that these samples were confirmed to show a single site of cleavage as confirmed by Edman sequencing of the postenzymology solid (glass-bound) sample and by amino acid analysis and LC/MS of the supernatant. Two columns are shaded differently to highlight the groups which show activity greater than that of the reference, Tyr.

AGPST-OH. These data confirm that these substrates are processed by stromelysin between **T** (Thr) and **X** residues. This implies that the groups shown in Figure 5 interact at the S_1' site of the stromelysin active site. Thus, these data also provide some additional SAR type information. For example, the data show a preference of oxygen vs sulfur (by comparison of Ser(Bn) vs Cys(Bn)) and indicate β-branching to be deleterious (comparison of Ser(Bn) and Thr(Bn)). In conjunction with a 3D model of stromelysin, we have been able to rationalize this

Chart 1. Variants Examined in the Immobilized Peptides [COP-AGPST″X″T-(Acp)$_5$-βAla-AMP-CPG] as Potential Substrates for Stromelysin



data on the basis of the steric and electronic effects and have suggested additional novel P_1' groups.[36,37]

Conclusions

We have provided the first example to our knowledge for utilization of raw biological data to guide a genetic algorithm driven chemical synthesis. These algorithms further facilitate the process of lead discovery/optimization by reducing user bias. Exploration of stromelysin-selective P1′ residue information provides an example of how, based on the assay results from any given screen, one could obtain focused information by this approach. Employment of GA-based optimization would require synthesis of a small fraction[38] of the combinatorial population. This approach provides an alternative strategy to effectively explore diversity space without the construction and assay of large libraries to identify lead candidate(s). Genetic algorithms should provide a powerful tool to help focus drug discovery. We hope, that this first example of the demonstration of GA as a tool to guide chemical synthesis based problems would provide an incentive for further exploration of these tools

(35) The mass spectra of the two samples containing Ser(OBn) and hPhe residues were essentially identical except for the ratio of the major peaks to be roughly 2.7:1, which happens to be the relative activity of these two samples. LC/MS of a sample containing Tyr also gave a similar spectra and identical $(M + H)^+$ ions.

(36) Singh, J.; Ghose, A. Unpublished results.

(37) Translation of the stromelysin selective P1′ and P2′ information to a series of inhibitors using solid phase based combinatorial synthesis is obviously the next logical step. It is feasible to involve GA to facilitate in rapid resolution to the selective inhibitor identification/optimization problem.

for increasingly important chemical diversity based problems. The use of GA could easily be tailored to small molecule based diversity solutions.[39,40] This strategy/approach should be equally applicable to both lead identification and (or perhaps more suited for) lead optimization processes.

Note Added in Proof: Subsequent to the submission of this manuscript, a report describing the use of a genetic algorithm for lead generation has appeared (Weber et al. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2280−2282).

Acknowledgment. We would like to thank Drs. James Gainor and John Mallamo for helpful discussions and Mr. George Gonyea for LC/MS experiments.

Supporting Information Available: Listing of sequences and their normalized data for screening versus mSl-t (5 pages). This material is contained in many libraries on microfiche, immediately follows this article in the microfiche version of the journal, can be ordered from the ACS, and can be downloaded from the Internet; see any current masthead page for ordering information and Internet access instructions.

JA953172I

(38) Computationally it has been shown that, in general, GA should converge in about 10−15 generations (see ref 8a). This would mean that one would have to synthesize a total of ~600 samples, i.e. <0.002% of the combinatorial population of 3 200 000. One could easily synthesize these relatively small numbers of compounds as individual compounds (via parallel synthesis) and would not have to resort to preparing mixtures. This in turn would obviate considerations of alternate decoding strategies to identify active compounds.

(39) The version of the GA described here is applicable to an oligomeric based template: A_n-B_m-C_p-D_q or a non-oligomeric template where A_n, B_m, C_p, and D_q represent pendants on a core scaffold or a hybrid of these. Here, *n*, *m*, *p*, and *q* represent number of variables at the respective positions. These templates may represent peptidomimetic, peptoid, or small molecule based templates for lead selection/optimization either as protease inhibitors or for receptor antagonists.

(40) The templates recently described by Martin et al. (ref 3 above) and Kick and Ellman (Kick, E. K.; Ellman, J. A. *J. Med. Chem.* **1995**, *38*, 1427−1430) are the two relevant examples to which the current version of the genetic algorithms may be potentially applied for the identification/optimization of lead compounds, respectively.