Journal of Computer-Aided Molecular Design, **14:** 487–494, 2000.
KLUWER/ESCOM

De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks

Gisbert Schneider*, Man-Ling Lee, Martin Stahl & Petra Schneider
F. Hoffmann-La Roche Ltd, Pharmaceuticals Division, CH-4070 Basel, Switzerland

Received 26 June 1999; Accepted 7 January 2000

Key words: de novo design, diversity, docking, genetic algorithm, peptide mimetic, pharmacophore, similarity, thrombin, virtual screening

Summary

An evolutionary algorithm was developed for fragment-based de novo design of molecules (TOPAS, *TOP*ology-*A*ssigning *S*ystem). This stochastic method aims at generating a novel molecular structure mimicking a template structure. A set of ~25,000 fragment structures serves as the building block supply, which were obtained by a straightforward fragmentation procedure applied to 36,000 known drugs. Eleven reaction schemes were implemented for both fragmentation and building block assembly. This combination of drug-derived building blocks and a restricted set of reaction schemes proved to be a key for the automatic development of novel, synthetically tractable structures. In a cyclic optimization process, molecular architectures were generated from a parent structure by virtual synthesis, and the best structure of a generation was selected as the parent for the subsequent TOPAS cycle. Similarity measures were used to define 'fitness', based on 2D-structural similarity or topological pharmacophore distance between the template molecule and the variants. The concept of varying library 'diversity' during a design process was consequently implemented by using adaptive variant distributions. The efficiency of the design algorithm was demonstrated for the de novo construction of potential thrombin inhibitors mimicking peptide and non-peptide template structures.

Introduction

Automated de novo design of bioactive molecules is one of the aspired goals in computational chemistry [1, 2]. To approach this idea, algorithms must be provided addressing two principal tasks: First, the search space, i.e., the set of all algorithmically tractable molecules, must be structured into regions of higher and lower quality to allow for prediction of desired properties (e.g., receptor binding). Second, a systematic search strategy must be at hand to facilitate navigation in a high-dimensional chemical space. Evolutionary algorithms have proven their value in a number of such 'virtual screening' experiments, and some examples of successful de novo designs have been published [3–6]. Here we present an algorithmic solution to the problem of evolutionary, template-based de novo design (TOPAS, *TOP*ology-*A*ssigning *S*ystem). Novel molecular compounds are suggested in a fully automated cyclic process, taking a given structure as a reference point ('seed' or 'template' structure). Furthermore, instead of generating molecular architectures containing undesired structural features, or synthetically intractable compounds – a problem encountered by many de novo design procedures [7] – TOPAS was equipped with a limited set of drug-derived building blocks, which were obtained from straightforward retro-synthetic fragmentation of the World Drug Index (WDI, Derwent Information, London). A strategy very similar to the RECAP procedure developed by Hann and co-workers was applied [8]. The idea is that re-assembly of such pre-defined building blocks by a limited set of chemical reactions might lead to chemically feasible novel structures. The follow-

*To whom correspondence should be addressed. E-mail: gisbert.schneider@roche.com ing additional objectives guided the development of TOPAS:

- Identification of 'fast-followers' taking a known lead or drug as the template structure.
- Generation of focused libraries that are biased towards a given activity for secondary screening.
- Preferred generation of novel 'drug-like' molecules for combinatorial library design.
- Development of peptide-derived molecules with a non-peptide backbone architecture.

In this work, TOPAS was used to breed potential thrombin inhibitors taking either the known thrombin inhibitor NAPAP (N-alpha-(2-naphthyl-sulphonyl-glycyl)-DL-p-amidinophenylalanyl-piperidine) [9, 10], or a natural peptidic thrombin substrate (Phe-Pro-Arg) as the template structure [11].

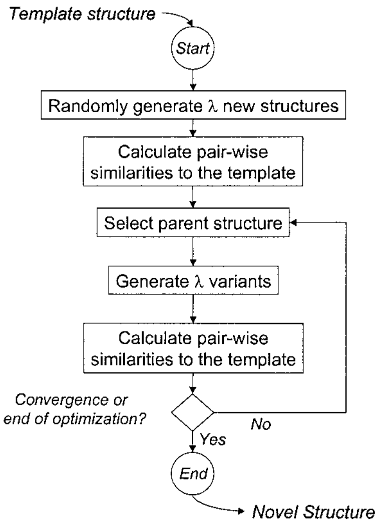

Template structure
Start
Randomly generate λ new structures
Calculate pair-wise similarities to the template
Select parent structure
Generate λ variants
Calculate pair-wise similarities to the template
Convergence or end of optimization?
No
Yes
End
Novel Structure

Figure 1. Scheme of the TOPAS design algorithm.

Methods

Fragmentation of drug molecules

All 36,000 structures contained in the WDI (as of November 1998, Derwent Information, London) that had an entry related to 'mechanism' or 'activity' were subjected to retro-synthetic fragmentation to compile a stock of building blocks for evolutionary de novo design by TOPAS. The reactions listed in Table 1 were applied to exhaustive cleavage with the following restrictions:

- Ring systems were not destroyed.
- Bonds between hetero atoms and ring carbons were not cleaved.
- If a terminal group was hydrogen, methyl, ethyl, propyl, or butyl, the reactions were not applied.

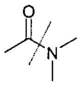
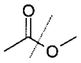
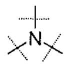
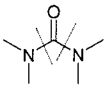
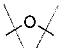
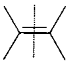
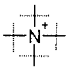
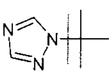
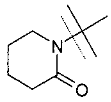
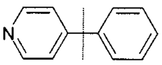
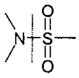
This procedure led to a total number of 24,563 unique building blocks for TOPAS ('stock of structures'). The list of fragmentation schemes is identical to the original RECAP procedure (Table 1) [8].

The TOPAS design algorithm

TOPAS is based on a simple evolutionary algorithm (EA), a (1,λ) evolution strategy [12]. The choice of a specific EA is almost arbitrary provided that the optimization process is guided by adaptive 'strategy parameters' (see below). The underlying principle is grounded on a stochastic search process (Figure 1): Starting from an arbitrary point in search space, the so-called initial 'parent', a set of λ 'variants' are generated. In this work we used $\lambda = 100$. A bell-shaped distribution of variants is generated for each 'generation' (optimization cycle), centered at the parent. This means that most of the variants are very similar to their parent structure, and with increasing dissimilarity (distance) the number of variants decreases. A 'fitness' value was calculated for each variant, and the 'fittest' one was selected as the parent of the subsequent generation. In the present study, this cyclic process was forced to terminate after 100 generations, which proved to be sufficient for convergence on a local fitness optimum. The process was assumed to have reached a fitness optimum when (i) no successful structural modification occurred during the past ten generations, and (ii) the strategy parameter 'step-size' approached a value of 0 (see below).

Fitness was defined as pair-wise similarity between the template and a new variant structure. Two different concepts were realized to measure similarity: (i) 2D structural similarity as defined by the Tanimoto index on Daylight's 2D fingerprints (for a review, see [13]), (ii) 2D topological pharmacophore similarity [14, 15]. Tanimoto similarity varies between 0 and 1, where the value of 1 indicates structural identity. Topological pharmacophore similarity values vary between 0 (indicating identical pharmacophore distribution in the two molecules) and positive values indicating varying degrees of pharmacophore similarity. Optimal fitness values are 1 for the Tanimoto measure, and 0 for the pharmacophore similarity measure.

Table 1. Fragmentation types used for fragmentation of the WDI compounds [8], and for de novo assembly of building blocks by TOPAS. The total number of fragments is 24,563. Dotted lines indicate the bond cleavage positions

Bond or structure type	Fragmentation type	No. of fragments in stock
Amide		5244
Ester		2257
Amine		4495
Urea		743
Ether		3101
Olefin		4300
Quarternary nitrogen		321
Aromatic nitrogen – aliphatic carbon		1637
Lactam nitrogen – aliphatic carbon		181
Aromatic carbon – aromatic carbon		1615
Sulphonamide		669

The general idea of our topological pharmacophore representation is to measure distances between pairs of atoms and regard the histogram of pair counts (two-point pharmacophores) as a simplifying but exhaustive pharmacophore fingerprint of the molecule. Distances are expressed as the number of bonds along the shortest path connecting two nodes (non-hydrogen atoms) in the molecular graph. Each node is checked as to whether it can be assigned one of the following generalized atom types: hydrogen-bond donor (D), hydrogen-bond acceptor (A), positively charged (P), negatively charged (N), or lipophilic (L). The numbers of all 15 possible pairs of generalized atom types (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) are determined, and the resulting histogram counts are divided by the total number of non-hydrogen atoms to obtain scaled vectors. Distances up to ten bonds showed to be relevant in most of the cases investigated until today, although the optimal path length varies in different applications (G. Schneider, unpublished). This leads to a 15×10 = 150-dimensional vector representation of a mole-

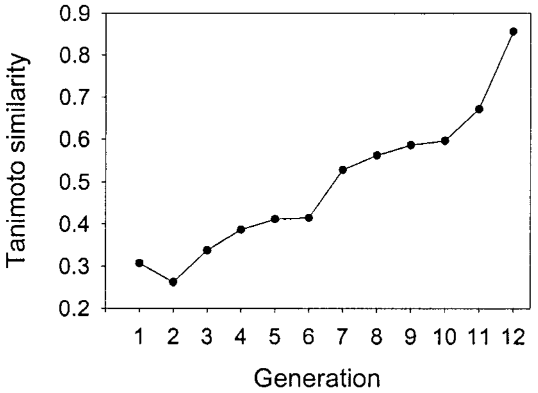


Figure 2. Fitness values of parent molecules of a TOPAS run. 'Fitness' was defined by the Tanimoto similarity between the designed molecules and the template (NAPAP).

cule. The Euclidian distance was calculated to express similarity between the template and novel structures [15].

Penalty terms were added to the fitness function to avoid undesired structures if the total number of non-hydrogen atoms exceeded 50, the sum of oxygen and nitrogen atoms was greater than 12, or if there were more than seven potential hydrogen-bond donors present in a given molecule. These criteria are thought to reduce the risk of designing molecules showing poor absorption or membrane permeation properties. They represent simple rules of thumb for restricting the search space to regions possibly enriched in 'drug-like' molecules [16].

The optimization employed 'adaptive step-size control', a strategy parameter which was implemented to enable a search process in chemical space that automatically adjusts to the local shape of the fitness landscape [12, 17]. This was achieved by subjecting the variance, σ, of the variant distribution to optimization. The algorithm can be written in pseudo-code:

```
initialize parent: ($S_P$, $\sigma_P$, $F_P$)
for each generation:
  generate $\lambda$ variants, ($S_V$, $\sigma_V$, $F_V$)
  select best variant structure,
    ($S_{best}$, $\sigma_{best}$, $F_{best}$)
  set ($S_P$, $\sigma_P$, $F_P$) = ($S_{best}$, $\sigma_{best}$, $F_{best}$)
```

Here, S represents the chemical structure, σ is the variance (step-size of the evolutionary optimization process), and F is the fitness value associated with S. Indices P and V refer to parent or variant attributes, respectively. In the first generation σ_P was set to unity, and S_P was a randomly assembled molecular structure. This was performed by first randomly picking one building block (the core fragment) from the stock of structures, and then adding further building blocks to all potential attachment sites of the core. The attachment sites of the core fragment and the appropriate reactions for building block attachment were defined by fragmentation of the WDI (see previous section, and Table 1). In each of the following generations novel variant structures, S_V, and the associated step-sizes, σ_V, were generated using the Box–Muller formula for calculation of Gaussian-distributed random numbers, g (where i and j are random numbers in]0,1]):

$$g = \sqrt{-2\ln(i)}\,\sin(2\pi j).$$
$$\sigma_V = \sigma_P + g.$$

Variant structures were derived from the parent molecule in a four-step process: (i) retro-synthetic fragmentation of S_P (Table 1), (ii) random selection of one of the generated fragments, (iii) substitution of this fragment by the one from the stock of building blocks having the pair-wise similarity index closest to the random number g, (iv) virtual synthesis to assemble the novel chemical structure. In the present implementation only fragments from the same fragmentation type can be substituted in step (iii), e.g., only stock fragments from the amide-cleaved group can be used as substitutes after retro-synthetic amide cleavage of the parent molecule. Furthermore, there is no special operator in the EA accounting for reducing or increasing the number of fragments per molecule. It turned out that large variations in compound size may occur during a design run as a consequence of the large number of different stock fragments. The number of fragments in a molecule is not fixed during optimization due to the selection process in step (ii) (see for example Figure 3): In each generation there is a chance to grow new molecules from a small core structure, which happens if a large part of a molecule is substituted by a tiny fragment of the same fragmentation type (Figure 3, step 2).

TOPAS was implemented as a set of C-modules including libraries of the Daylight Toolkit (C.A. James & D. Weininger, Daylight Chemical Information Systems Inc., Irvine, CA).

Results and discussion

As a first application of TOPAS we tried to develop molecules mimicking the NAPAP structure. The Tan-

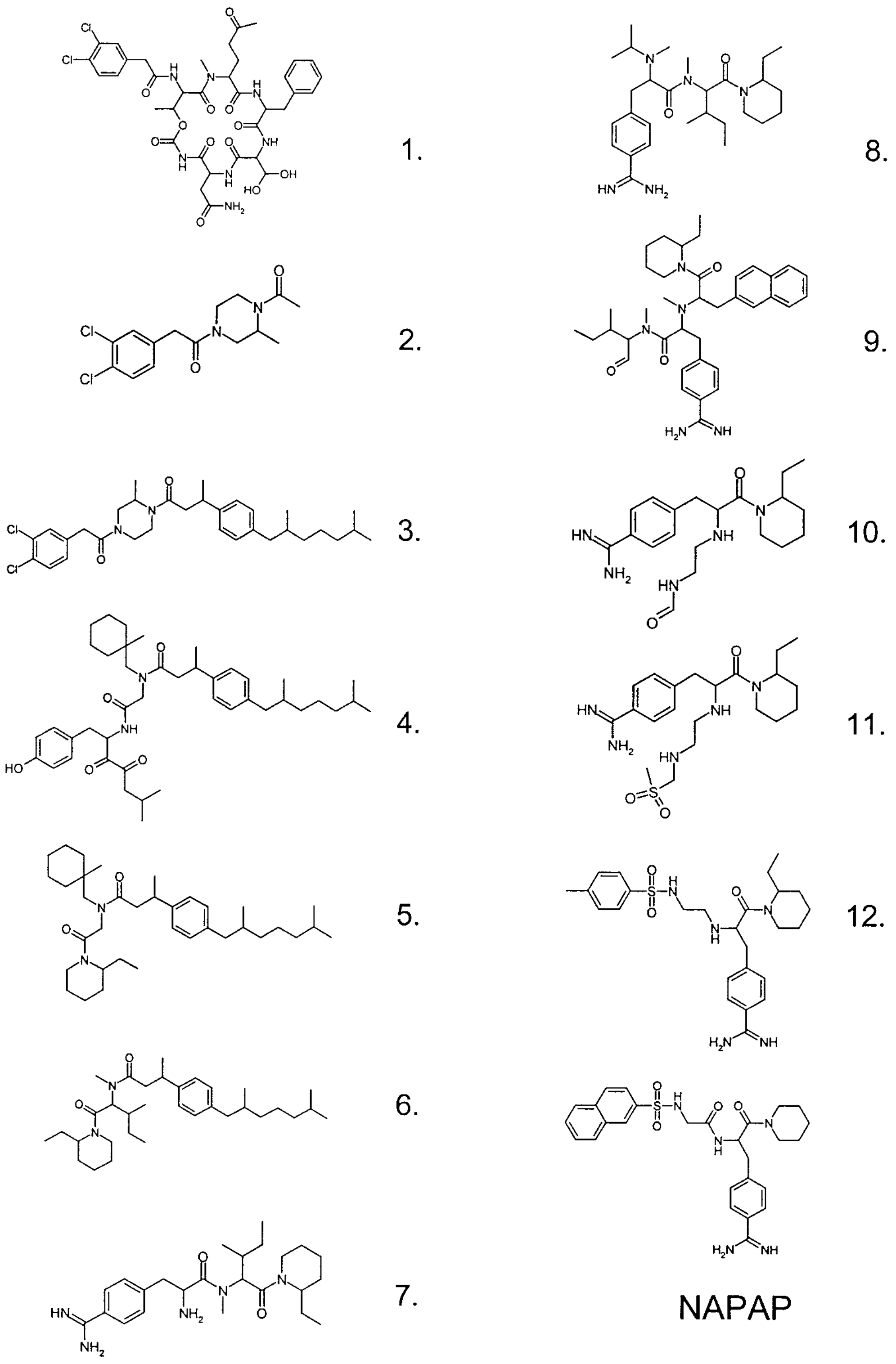


Figure 3. Parent molecules of 12 generations of a TOPAS run. The task was to design a molecule mimicking the NAPAP structure. ‘Fitness’ was defined by the Tanimoto similarity between the designed molecules and the template (NAPAP).

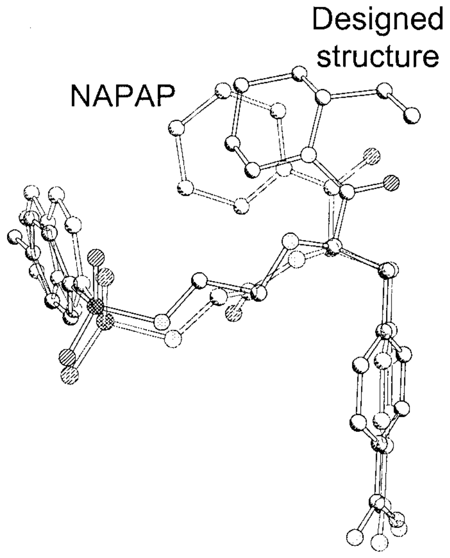


Figure 4. Structural models of NAPAP (gray, 1dwd X-ray structure) and a TOPAS-designed molecule (black, FlexX structure) in the thrombin active site.

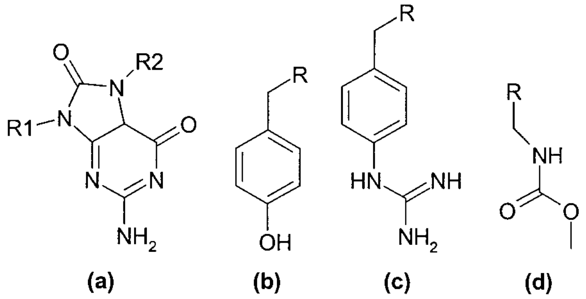


Figure 5. 'Needle' residues found by TOPAS. These structures are thought to mimic the NAPAP benzamidine moiety. R denotes the residue attachment site.

imoto index was used as the fitness measure, and the aim was to systematically assemble molecular building blocks in a way that the novel structures have a high pair-wise Tanimoto similarity to NAPAP. Initially, a random assembly of molecular fragments picked from the stock of 24,000 available building blocks was performed ('parent' of the first generation). The Tanimoto similarity to NAPAP was 0.31, reflecting a high dissimilarity, as expected. In each of the following generations 100 variants were systematically generated by TOPAS, and the best of each generation was selected as the parent for the subsequent generation (Figure 1). Following this scheme, novel molecules were assembled which exhibited a significantly increased 'fitness' (Figure 2). After only 12 optimization cycles the process converged at a high fitness level (approximately 0.86). This is to be expected due to the nature of the similarity measure [13]. At this stage the optimization process was stopped because the strategy parameter σ approached the value 0, and no further increase in fitness (Tanimoto index) was observed. The parent structures of each generation are shown in Figure 3. The resulting final design indeed shares a significant set of substructure elements with the template (NAPAP). Essential key features for thrombin binding evolved: the benzamidine moiety forming hydrogen bonds with Asp189, a sulfonamide moiety interacting with the backbone carbonyl of Gly216, and the lipophilic para-tolyl and piperidine rings filling a large lipophilic pocket of the thrombin active site cleft (for a more detailed description of the thrombin active site region and specific inhibitor interactions, see e.g. References 18 and 19 and literature cited therein). Automated docking by means of FlexX [20] essentially reproduced the NAPAP binding mode (PDB code 1dwd [21]), substantiating our conclusions (Figure 4). However, this observation alone does not prove that all TOPAS designs will be compatible to the steric and electrostatic properties of the binding pocket, since it is well known that binding modes obtained by flexible docking (e.g., by FlexX) are very sensitive to small changes in the ligand structure. In general the probability to obtain identical binding modes for different molecular backbone architectures is inversely proportional to the probability for finding a novel scaffold. The latter mainly depends on the restrictions dictated by the similarity (fitness) measure.

This first 'from scratch' TOPAS design experiment clearly demonstrated that the algorithm can be used for a fast guided search in a very large chemical space, ending up with rational proposals for novel molecular structures that are similar to a given template. We repeated the identical experiment several times, ending up with slightly different designs. This indicates that TOPAS is easily trapped in a local optimum, which very likely is a consequence of the very coarse-grained definition of building blocks and the restricted set of reactions allowed. Fine-tuning of structures can hardly be expected from the current implementation.

In the past, moieties binding to the thrombin P1 pocket were subjected to many modification attempts, with the twofold aim to enhance binding specificity of the inhibitor and avoid troublesome properties of the arginine side chain [19]. We tested TOPAS for its ability to abstract from the NAPAP benzamidine residue and suggest potential substitutes. In this ex-

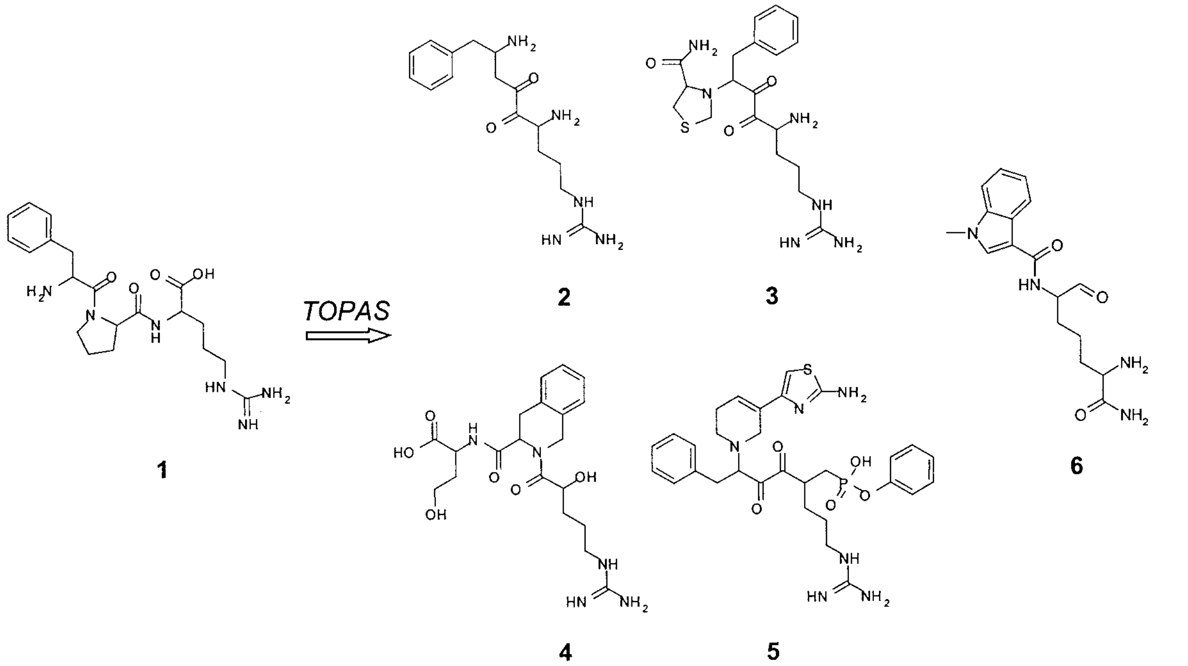


Figure 6. Peptide-mimicking designs obtained from five independent TOPAS runs. The task was to generate novel structures taking the tripeptide Phe-Pro-Arg as the template.

periment the pharmacophore similarity measure was employed instead of the Tanimoto structural similarity measure. In Figure 5 some results are shown (structures (a)–(d)). We have not found any experimental support for needle (a) binding to the P1 pocket. Although this purine derivative has an appealing H-bond donor/acceptor pattern, the carbonyl might be detrimental to P1 pocket binding. However, the Roche corporate database comprises several potent thrombin inhibitors containing needles (b), (c), and (d) (K_i in the nanomolar and low micromolar range, structures not shown). Several potent thrombin inhibitors with neutral H-bond donating phenols (b) and structures similar to (d) were also reported by others [19]. It is evident that TOPAS was able to derive interesting alternatives to benzamidine. Most critical for the design was the appropriate selection of a similarity measure as the fitness function. The pharmacophore similarity as defined in TOPAS seems to represent a reasonable choice. Recently, this similarity measure was successfully applied to 'backbone-hopping' between calcium antagonists, i.e., identification of isofunctional molecules with a significantly different architecture [15].

A further application of TOPAS was the design of peptide-analogues taking the tripeptide D-Phe-Pro-Arg – an excellent natural thrombin substrate [11] – as the template structure. Five independent TOPAS experiments were performed. Each run was forced to stop after 100 generations with 100 variants per generation tested. The topological pharmacophore similarity was used to measure fitness to find structures exhibiting a distribution of functional groups that is

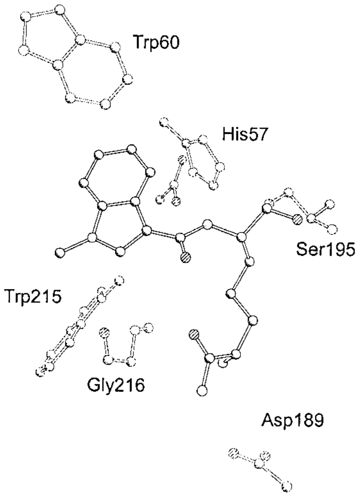


Figure 7. Structural model of a peptide-derived molecule (design **6**) in the thrombin active site.

similar to the peptidic template, yet with a non-peptidic backbone architecture. The design processes rapidly led to high-fitness structures ending up with surprising results (Figure 6). All five designs (structures **2**–**6**) contain a reasonable P1 needle. In the structures **2**, **3**, **4**, and **5** the original arginine side-chain was selected by TOPAS, whereas structure **6** contains an interesting alternative. Furthermore, lipophilic moieties are present in appropriate positions, possibly filling lipophilic pockets of the thrombin active site. Manual docking of structure **6** into the thrombin active site (PDB code 1dwd) and subsequent energy minimization using the MAB force field of the MOLOC package [22] led to a reasonable model (Figure 7). Although several peptidic features are still present in the peptide-derived structures, a number of surprising alternatives were found. TOPAS clearly demonstrated its capability to evolve well-known features of small-molecule thrombin inhibitors from a peptidic structure, e.g., the classical arginine aldehyde pattern for covalent binding to the catalytic Ser195 (like in efegatran) or the α-ketoamide derivative (like in CVS 863) [19]. It must be stressed that the topological pharmacophore similarity employed here does not distinguish between backbone and side chain functionality. Therefore, without appropriate weighing of individual contributions of functional groups one cannot expect bioactive designs from each design experiment. Still the medicinal chemist must evaluate each novel structure. The main aim of TOPAS is to make suggestions for potential make-ups of novel molecular architectures.

Some TOPAS designs have already been synthesized and tested at F. Hoffmann-La Roche Ltd, Basel. Some of the suggested molecules exhibit substantial bioactivity (not shown, unpublished). Although there is ‘proof of concept’ already, TOPAS represents an approach in its early stages. Its main limitation is the inability to perform fine-optimization. This restriction could be overcome by enlarging the set of possible reactions for the later optimization stages, or by excessively increasing the stock of building blocks. An additional option is to modify the design algorithm in a way that multiple fragments may be substituted within each generation. Irrespective of the outcome of the biochemical activity assays the proposals generated by TOPAS can help the medicinal chemist to derive hypotheses about structure–activity relationships, and guide the required synthetic work. Furthermore, the proposed novel structures are excellently suited for subsequent evaluation by structure-based modeling, virtual screening, and other property prediction techniques [7, 23].

Acknowledgements

Gérard Schmid, Christian Hubschwerlen, Werner Neidhart, Alexander Alanine, Daniel Doran, Geo Adam, and Hans-Joachim Böhm (all F. Hoffmann-La Roche Ltd, Basel) are thanked for inspiring discussions and encouragement.

References

1. Böhm, H.-J., J. Comput.-Aided Mol. Design, 12 (1998) 309.
2. Kubinyi, H., J. Recept. Signal Transduct. Res., 19 (1999) 15.
3. Willett, P., Trends Biotechnol., 13 (1995) 516.
4. Schneider, G., Schrödl, W., Wallukat, G., Nissen, E., Rönspeck, G., Müller, J., Wrede, P. and Kunze, R., Proc. Natl. Acad. Sci. USA, 95 (1998) 12179.
5. Walters, W.P., Stahl, M.T. and Murcko, M.A., Drug Discov. Today, 3 (1998) 160.
6. Wrede, P., Landt, O., Klages, S., Fatemi, A., Hahn, U. and Schneider, G., Biochemistry, 37 (1998) 3588.
7. Böhm, H.-J., Banner, D.W. and Weber, L., J. Comput.-Aided Mol. Design, 13 (1999) 51.
8. Lewell, X.Q., Judd, D.B., Watson, S.P. and Hann, M.M., J. Chem. Inf. Comput. Sci., 38 (1998) 511.
9. Kaiser, B., Hauptmann, J., Weiss, A. and Markwardt, F., Biomed. Biochim. Acta, 44 (1985) 1201.
10. Bode, W., Turk, D. and Sturzebecher, J., Eur. J. Biochem., 193 (1990) 175.
11. Mohler, M.A., Refino, C.J., Chen, S.A., Chen, A.B. and Hotchkiss, A.J., Thromb. Haemost., 56 (1986) 160.
12. Rechenberg, I., Evolutionsstrategie – Optimierung technischer Systeme nach Prinzipien der biologischen Evolution (extended edition 1994), Frommann-Holzboog, Stuttgart, 1973.
13. Willett, P., Barnard, J.M. and Downs, G.M., J. Chem. Inf. Comput. Sci., 38 (1998) 983.
14. Schneider, G. and Wrede, P., Prog. Biophys. Mol. Biol., 70 (1998) 175.
15. Schneider, G., Neidhart, W., Giller, T. and Schmid, G., Angew. Chem. Int. Ed. Engl., 38 (1999) 2894.
16. Walters, W.P., Ajay and Murcko, M.A., Curr. Opin. Chem. Biol., 3 (1999) 384.
17. Schneider, G., Schuchhardt, J. and Wrede, P., Biol. Cybern., 74 (1996) 203.
18. Grootenhuis, P.D. and Karplus, M., J. Comput.-Aided Mol. Design, 10 (1996) 1.
19. Wiley, M.R. and Fisher, M.J., Exp. Opin. Ther. Patents, 7 (1997) 1265.
20. Rarey, M., Wefing, S. and Lengauer, T., J. Comput.-Aided Mol. Design, 10 (1996) 41.
21. Banner, D.W. and Hadváry, P., Biol. Chem., 266 (1991) 20085.
22. Gerber, P.R. and Müller, K., J. Comput.-Aided Mol. Design, 9 (1995) 251.
23. Bohacek, R.S. and McMartin, C., Curr. Opin. Chem. Biol., 1 (1997) 157.