INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0418

A FAST EXCHANGE ALGORITHM FOR DESIGNING FOCUSED LIBRARIES IN LEAD OPTIMISATION

C. LE BAILLY de TILLEGHEM, B. BECK, B. BOULANGER, and B. GOVAERTS,

http://www.stat.ucl.ac.be

A Fast Exchange Algorithm for Designing Focused Libraries in Lead Optimisation

C. Le Bailly de Tilleghem[1] B. Beck[2] B. Boulanger[2] B. Govaerts[1]

Abstract

Combinatorial chemistry is widely used in drug discovery. Once a lead compound has been identified, a series of R-groups and reagents can be selected and combined to generate new potential drugs. The combinatorial nature of this problem leads to chemical libraries containing usually a very large number of virtual compounds, far too large to permit their chemical synthesis. Therefore, one often wants to select a subset of "good" reagents for each R-group of reagents and synthesise all their possible combinations. In this research, one encounters some difficulties.

First, the selection of reagents has to be done such that the compounds of the resulting sub-library simultaneously optimise a series of chemical properties. For each compound, we use a desirability index, a concept proposed by Harrington [20], to summarise those properties in one fitness value. Then a loss function is used as objective criteria to globally quantify the quality of a sub-library.

Secondly, there are a huge number of possible sub-libraries and the solutions space has to be explored as fast as possible. The *WEALD* algorithm proposed in this paper starts with a random solution and iterates by applying exchanges, a simple method proposed by Federov [13] and often used in the generation of optimal designs. Those exchanges are guided by a weighting of the reagents adapted recursively as the solutions space is explored.

The algorithm is applied on a real database and reveals to converge rapidly. It is compared to results given by two other algorithms presented in the combinatorial chemistry literature: the *Piccolo* algorithm of W. Zheng et al. [37] and the *Ultrafast* algorithm of D. Agrafiotis and V. Lobanov [4].

Keywords: drug discovery, combinatorial chemistry, hit-to-lead, library design, reagents selection, multiobjective optimisation, desirability.

1 Introduction

The process of drug discovery generally implies the following successive stages. Once a biological target has been defined, the screening of large collections of compounds is performed

---

[1] Institute of Statistics from the Université catholique de Louvain – 20, voie du roman pays, 1348 Louvain-la-Neuve, Belgium – lebailly@stat.ucl.ac.be

[2] Lilly Services S.A. – 11, rue Granbonpré, 1348 Mont-Saint-Guibert, Belgium

to identify *hits* i.e. compounds that interact with the biological target. Then *leads* are generated by chemically modifying the hits to improve chemical properties. Further chemical alterations of leads are performed to optimise criteria and convert leads into drug development candidates. Finally, preclinical trials confirm the characteristics of the new potential drug. Those different stages are summarised in Figure 1. See [26, 27, 28, 9, 10, 12] for a review on drug discovery and development.

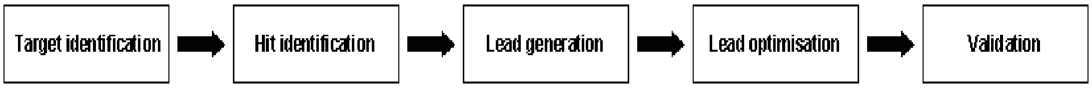


Figure 1: General scheme of drug discovery process.

Combinatorial chemistry is widely used in this process [15, 19, 14, 23, 6, 24, 35]. This technique systematically combine a variety of molecular building blocks into a huge collection of compounds called *combinatorial library.* There is an increasing need of methods to design combinatorial libraries [32, 33, 29]. A first possible objective is to span a wide range of chemical properties. This is known under the name *diverse library* or *exploratory library* [3, 4, 37]. It can be used in the hit identification stage, when chemists want to select active compounds as diverse as possible. The design of libraries using some diversity measurements is often discussed in literature [17, 16, 21, 2, 8, 5, 1, 25]. A second possible objective is to obtain a library biased toward a specific target or optimising some chemical properties. This is known under the name *focused library* or *targeted library* [3, 4, 37]. It can be used in the lead optimisation stage.

There is an increasing need of methods to construct combinatorial libraries. Part of those techniques are borrowed from the field of operational research, based on simulated annealing algorithms or genetic algorithms [37, 30, 34, 31, 36, 18, 7, 22]. This paper presents a new method based on exchanges to design focused library in lead optimisation. This iterative algorithm finds rapidly in the combinatorial library a sub-library of reasonable size that is composed of promising molecules that could be easily synthesised in laboratories thanks to their combinatorial nature.

Some notations are first fixed and the general aim of focused library design is defined using a real combinatorial library provided by Eli Lilly S.A. as an example. Then the principles of the *WEALD* algorithm are explained and the results of its application on the combinatorial library shows how well it performs. Finally, the *WEALD* algorithm is compared with two other existing methods, the *Ultrafast* algorithm of Agrafiotis and Lobanov [4] and the *Piccolo* algorithm of Zengh et al. [37].

2 Designing focused library in lead optimisation

Once a lead has been generated, chemists in pharmaceutical industries try to chemically modify the lead to obtain compounds having still better properties. They virtually divide the lead into N basic parts and set up a list of possible modifications on each part. Each modification is called a *reagent* (or *reactant*) and each set of modifications on a part of the lead is called a *R-group* (or a *pool*) [3, 4, 37].

Let N be the number of R-groups and N_i the number of reagents in the i^{th} R-group ($i = 1, 2, \ldots, N$). The number of possible compounds obtained by combining a reagent from each R-group is $M = \prod_{i=1}^{N} N_i$. They form the combinatorial library. In most cases, its size is so huge that it is impossible to synthesise all the compounds in the library. Chemists have thus to select in each R-group n_i reagents ($n_i \leq N_i$, $i = 1, 2, \ldots, N$) and combine them in a full fashion to obtain a sub-library of reasonable size. The number of compounds in the resulting sub-library is $m = \prod_{i=1}^{N} n_i$. This more reasonable number and the combinatorial nature of the resulting compounds allow an easy synthesis in laboratories.

In the lead optimisation stage, reagents have to be selected in order to get the *best* sub-library, i.e. a sub-library containing compounds that optimise some chemical criteria. As there are $\binom{N_i}{n_i}$ different ways to select n_i reagents in the i^{th} R-group, the number of all possible sub-libraries is $\prod_{i=1}^{N} \binom{N_i}{n_i}$, which is often huge in practice.

Along the next sections, a real combinatorial library is used to show the performance of the *WEALD* algorithm for optimising sub-libraries and to compare it to other methods. This library has been developed in the context of a new serotoninergic antidepressant and is referred as the *antidepressant library.*

The library is composed of $N = 4$ R-groups. The different R-groups are represented on Figure 2. There are $N_1 = 12$, $N_2 = 19$, $N_3 = 12$ and $N_4 = 41$ reagents in each R-group. The size of the whole combinatorial library is thus $M = \prod_{i=1}^{N} N_i = 12 \times 19 \times 12 \times 41 = 112176$, which is far too much to test all those compounds in vitro.

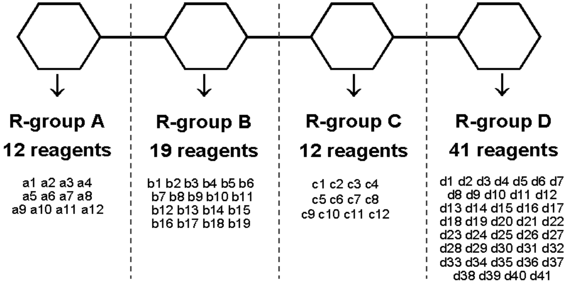


Figure 2: *antidepressant library*: real combinatorial library composed of 4 R-groups and $12 \times 19 \times 12 \times 41 = 112176$ compounds.

Chemists at Eli Lilly want to select $n_i = 3$ $(i = 1, 2, \ldots, N)$ reagents in each R-group to obtain a sub-library of $m = \prod_{i=1}^{N} n_i = 3^4 = 81$ compounds that optimises 8 criteria of interest (e.g. efficiency, binding, absorption, distribution, metabolism, excretion, toxicity,...). Those criteria are predicted according to chemical descriptors (e.g. molecular weight, number of bounds, number of hydrogen atoms...). As there are $\prod_{i=1}^{N} \binom{N_i}{n_i} = \binom{12}{3} \times \binom{19}{3} \times \binom{12}{3} \times \binom{41}{3} \simeq 5 \times 10^{11}$ different ways to select three reagents in each R-group and combine them, there is a clear need for a method to find the one that optimises the 8 chemical criteria.

3 A Weighting Exchange Algorithm for Library Design (*WEALD*)

3.1 Principles

This paper proposes a new method to select reagents in R-groups to obtain compounds that optimise some properties. The algorithm is iterative as most of the algorithms in this domain.

Firstly, n_i reagents are selected at random in the i^{th} R-group and combined to form an initial sub-library with $m = \prod_{i=1}^{N} n_i$ compounds. Then exchanges are performed on the current sub-library to improve the quality of its compounds. The principle of exchanges was introduced in experimental D-optimal design by Federov in 1972 [13]. This principle is applied to select a reagent in the current sub-library and replace it by a new one. But exchanges are not done in a naive way at random. First a criteria has been developed to quantify in one index the quality of a sub-library. This is explained in section 3.2. Secondly a weighting method summaries all the information accumulated till each iteration to guide the reagents exchanges. This is explained with more details further in section 3.3.

3.2 Quality of a sub-library

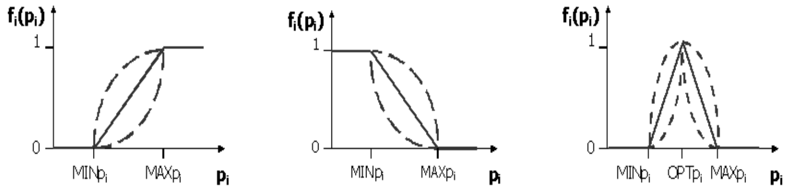
The algorithm objective is to select in a combinatorial fashion compounds that optimise 8 properties. In front of such multiobjective problem, the literature often proposes to aggregate those variables into a unique value between 0 and 1 that is a compromise of all the properties values. This is the *desirability* principle introduced by Harrington in 1965 [20]. The quality of a compound is quantified by a *fitness* value lying between 0 and 1 that summarises the properties of interest. Higher is the fitness better the compound fits the criteria.

Let P be the number of properties to optimise. In the *antidepressant library*, P equals 8 and the fitness of a molecule is computed using the Derringer principle [11] with the next formula:

$$Fitness = \prod_{i=1}^{P} \left[f_i\left(P_i\right)\right]^{w_i}$$

where P_i is the i^{th} property, f_i is a function that maps the i^{th} property in the interval $[0, 1]$ and w_i is a weight chosen to emphasise (or not) the i^{th} property with $\sum_{i=1}^{P} w_i = 1$.

Figure 3 displays three general forms of functions f_i proposed by Derringer. The first and the



s

Figure 3: Different kinds of functions f_i proposed by Derringer [11] to map chemical properties in the interval $[0, 1]$.

second graphs show increasing and decreasing functions that can be used when the property P_i has to be maximised or minimised respectively. The last graph shows a function f_i that can be used when the property P_i must reach an optimal value OPT_{P_i}.

The graphs of Figure 4 displays the values of the 8 properties to optimise computed for all the compounds in the combinatorial library. The best molecule is represented with the square on each graph and has a fitness value of 0.6908.

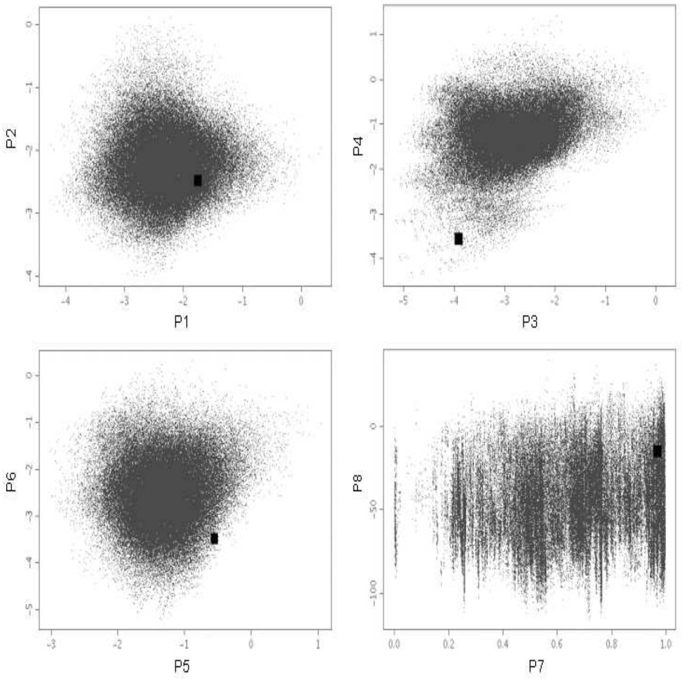


Figure 4: Values of the 8 properties $P1, P2, ..., P8$ computed for the 112176 compounds of the *antidepressant library*. The square represents the best molecule (fitness=0.6908)

The overall quality of a sub-library is quantified by a *loss* value lying between 0 and 1 that summarises the differences between the fitness of its m compounds and their ideal value 1:

$$Loss = \frac{1}{m}\sum_{j=1}^{m}[1 - Fitness_j]^2$$

Smaller is the loss better is the sub-library.

To summarise, a *fitness* value is associated to each compound (the highest, the best) and a *loss* value is associated to each sub-library (the smallest, the best). In the *WEALD* algorithm, exchanges between reagents are repeatedly performed to decrease the loss of the current sub-library.

3.3 Exchanges methodology

The *WEALD* algorithm starts by selecting at random n_i reagents in the i^{th} R-group and combine them to form an initial sub-library with $m = \prod_{i=1}^{N} n_i$ compounds. Then, repeatedly, exchanges are performed to obtain a better sub-library. In an exchange, one reagent is ejected from the current sub-library and replaced by another one of the same R-group to decrease the loss of the current sub-library.

First the reagent to be ejected out of the current sub-library is selected. For each current selected reagent, the mean of the fitness associated to the compounds in the sub-library containing this reagent is computed. The worst one is ejected, i.e. the one with the smallest average fitness.

In a second step, a new reagent is selected to replace the ejected one in the set of reagents of the same R-group that are not yet in the current sub-library. This choice is drown randomly according to probabilities of selection associated to each reagent. At the beginning of the algorithm, in each R-group, all reagents have the same probability to enter the sub-library. Then, at each iteration, probabilities of selection are adapted using all the fitness already computed according to the following rules:

Rule 1: If a reagent has never been tested for an exchange, it is associated to a selection probability of 1.

Rule 2: If a reagent has already been tested for an exchange, it is associated to a probability of selection which is the ratio between the average fitness of all the compounds already explored that contains this reagent and the maximum fitness computed till that step.

Rule 3: If the average fitness of the compounds obtained using a reagent is lower than the 10^{th} percentile of all the fitness computed till that step and if more than 1% of the compounds in the whole combinatorial library have been explored, the probability of selection for this reagent will be set at 0.

All these selection probabilities are of course standardised to sum to 1 in each R-group. Rule 1 insures that a reagent not explored yet will have the highest selection probability for next exchanges. With rule 2, exchanges with reagent that provides good compounds will be preferentially tested. And with rule 3 a bad reagent will definitely be excluded from the sub-library and never more tried for an exchange.

Finally, the exchange between the entering reagent drown randomly according to selection probabilities and the ejected worst one is applied if the loss of the new sub-library is lower than the loss of the current sub-library. If it is not the case, an other reagent of the same R-group is tried to replace the ejected one and if no exchange with the ejected reagent decreases the loss it is kept till the convergence of the algorithm.

The algorithm ends when there is no more possible exchange with reagent having selection probability strictly positive.

The different steps of the *WEALD* algorithm are summarised next:

Step 0: - Initialise a sub-library at random $S^0 = S_1^0 \times S_2^0 \times \ldots \times S_N^0$ where $S_i^0 = \{R_{ik}; k = 1, 2, \ldots, n_i\}$ is a set of selected reagents of the i^{th} R-group, R_i, and record its loss L^0;
- Set the best sub-library $S^b = S^0$ and its loss $L^b = L^0$;
- Initialise the selection probabilities for each reagent, R_{ij}, by $w_{ij} = \frac{1}{N_i}$ $(j = 1, \ldots, N_i$ and $i = 1, \ldots, N)$;
- Go to step 1.

Step 1: - If $w_{ij} = 0$ $\forall i$ and $\forall j$ stop and output S^b else execute step 2.

Step 2: - Select the worst reagent of S^b, denoted R_{IK};
- Select the set of possible reagents for an exchange, denoted $R^* = R_I S_I^b = \{R_{It}, t = 1, \ldots, N_i - n_i\}$;
- Go to step 3.

Step 3: - If $R^* \neq \emptyset$ and if $w_{It} \neq 0$ $\forall t$, generate a trial sub-library S^t by replacing R_{IK} in S^b by a reagent of R^*, denoted R_{IT}, chosen according to the selection probabilities; compute its loss L^t; go to step 4;
- If $R^* = \emptyset$ or if $w_{It} = 0$ $\forall t$, keep R_{IK} in S^b till convergence and never try it again for an exchange; update the selection probabilities; return to step 1.

Step 4: - If $L^t < L^b$ set $S^b = S^t$ and $L^b = L^t$; update the selection probabilities; return to step 1;
- If $L^t \geq L^b$ set $R^* = R^* R_{IT}$ and return to step 3;

3.4 Application on the *antidepressant library.*

We applied the *WEALD* algorithm (implemented in the S-Plus software) on the *antidepressant library* $(12 \times 19 \times 12 \times 41)$ to select $n_i = 3$ reagents in each R-group. By combining them, $m = 81$ compounds are obtained and they have to optimise 8 criteria of interest. As described in section 3.2, we use fitness values to quantify the quality of compounds and a loss value to measure the quality of a sub-library. The *WEALD* algorithm was applied 100 times, each time with a different random initial sub-library. This section analyses the rate of convergence of the *WEALD* algorithm and the quality of the sub-libraries obtained.

The first important result is that the *WEALD* algorithm converges rapidly. On average, 2613 fitness are computed before the algorithm stops. This represents only 2, 33% of the whole combinatorial library. In addition, this fast rate of convergence is quite stable, whatever is the initial sub-library. Indeed, each time the *WEALD* algorithm had been applied, the convergence was reached by computing between 2220 and 3114 fitness (between 2% and 2.8% of the whole library). This can be seen on the histogram (1) of Figure 5 representing the distribution of the numbers of fitness computations for the 100 runs of the algorithm.

The second important result is that the quality of the sub-libraries obtained with the *WEALD* algorithm is quite high. Indeed, losses of the 100 sub-libraries are quite small, on average 0.1382. They are also stable if we analyse the 100 losses represented on histogram (2) of Figure 5.

The fitness of the compounds in the 100 sub-libraries are high, on average 0.6291, which is the $99,71^{th}$ percentile of the fitness of the whole combinatorial library. One can conclude by analysing those numbers or comparing histograms (3) and (4) of Figure 5 that the *WEALD* algorithm provides a sub-library containing most of the best compounds (highest fitness) of the initial combinatorial library.

3.5 What happens when applying the *WEALD* algorithm without using selection probabilities?

We applied the principles of the *WEALD* algorithm on the *antidepressant library* but without using selection probabilities (algorithm called *Exchange Algorithm for Library Design - EALD*) to confirm that those proposed weights help finding more rapidly a sub-library 3^4 containing compounds that optimise the 8 properties of interest.

On 100 different initial sub-libraries, we applied both algorithms, *WEALD* and *EALD*, and compared the rates of convergence and the quality of the obtained sub-libraries. The results of those simulations are summarised on Figure 6. We conclude that using probabilities of selection accelerates the convergence because the *WEALD* algorithm only computes on average 2651 fitness which is only 2, 36% of the whole combinatorial although *EALD* computes on average 3553 fitness which corresponds to 3, 17% of the whole combinatorial library. Moreover, the *WEALD* algorithm provides slightly better sub-libraries than *EALD* according to the average losses that are smaller with *WEALD* (0.1372) than with *EALD* (0.1392) or according to the fitness of the selected compounds that are a little higher with *WEALD* (0.6304) than with *EALD* (0.6279).

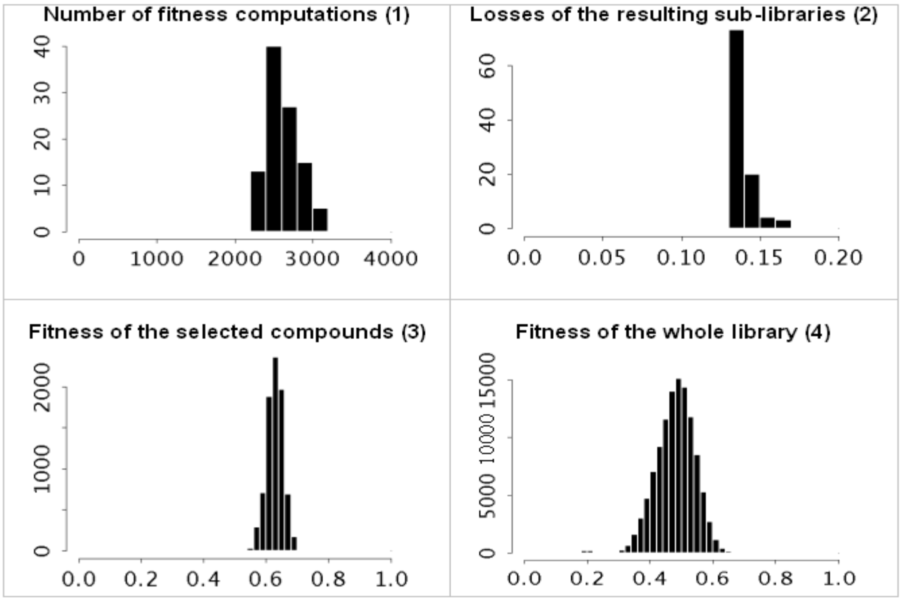


Figure 5: Results of the application of the *WEALD* algorithm on 100 different initial sub-libraries. (1) Number of fitness computations, (2) Losses of the final sub-libraries, (3) Fitness of the compounds in the final sub-libraries and (4) Fitness of the 112176 compounds in the whole library.

[redacted]