A POPULATION BASED APPROACH FOR HARD GLOBAL OPTIMIZATION PROBLEMS BASED ON DISSIMILARITY MEASURES

ANDREA GROSSO, MARCO LOCATELLI, AND FABIO SCHOEN

Abstract. When dealing with extremely hard global optimization problems, i.e. problems with a large number of variables and a huge number of local optima, heuristic procedures are the only possible choice. In this situation, lacking any possibility of guaranteeing global optimality for most problem instances, it is quite difficult to establish rules for discriminating among different algorithms. We think that in order to judge the quality of new global optimization methods, different criteria might be adopted like, e.g.:

(1) *efficiency* – measured in terms of the computational effort necessary to obtain the putative global optimum
(2) *robustness* – measured in terms of "percentage of successes", i.e. of the number of times the algorithm, re-started with different seeds or starting points, is able to end up at the putative global optimum
(3) *discovery capability* – measured in terms of the possibility that an algorithm discovers, for the first time a putative optimum for a given problem which is better than the best known up to now

Of course the third criterion cannot be considered as a compulsory one, as it might be the case that for a given problem the best known putative global optimum is indeed the global one, so that no algorithm will ever be able to discover a better one.

In this paper we present a computational framework based on a population-based stochastic method in which different candidate solutions for a single problem are maintained in a population which evolves in such a way as to guarantee a sufficient diversity among solutions. This diversity enforcement is obtained through the definition of a dissimilarity measure whose definition is dependent on the specific problem class. We show in the paper that, for some well known and particularly hard test classes the proposed method satisfies the above criteria, in that it is both much more efficient and robust when compared with other published approaches. Moreover, for the very hard problem of determining the minimum energy conformation of a cluster of particles which interact through short-range Morse potential, our approach was able to discover 4 new putative optima.

Keywords: large scale global optimization, multistart, basin hopping, population-based approaches, molecular conformation problems.

1. Introduction

We consider Global Optimization (GO) problems with the form

$$\text{minimize } f(X) \quad \text{subject to } X \in \mathcal{D} \tag{1}$$

where $\mathcal{D}$ is a box in $\mathbb{R}^n$ and f is a highly multimodal function. We also assume that efficient *local* optimization procedures exist for these problems. According to [14], the number of local minima of f over $\mathcal{D}$ is not — in general — an appropriate measure for the difficulty of (1): the latter also comes from the way local minima are organized on the landscape of f. A key concept to measure the difficulty, first introduced in global optimization problems arising in computational chemistry, is that of a *funnel*. In order to roughly describe what a funnel is, we can think of a graph whose nodes are local optima; two local optima X_i and X_j with $f(X_j) \leq f(X_i)$ are connected by a directed arc if from X_i it is possible to reach X_j. This possibility might be interpreted and defined in different ways. In chemistry and biology reachability corresponds to the situation in which there exists a continuous path connecting the two configurations which never exceeds a given energy level. So we might define as connected by an arc two local minima such that there is a path connecting them along which the objective function never exceeds a given value. Alternatively, we might say that X_j is reachable from X_i if a local optimization started from a point in a neighbor of X_i ends up at X_j. In any case, given a definition of reachability, a funnel bottom is defined as a local minimum with no outgoing arcs and a funnel is defined as a maximal set of local optima from which the same funnel bottom can be reached through a directed path.

Thus a funnel is a set of local minima characterized by the fact for each of them there exists at least one decreasing sequence of "neighbor" local minima along a path leading to a unique local minimum corresponding to the *bottom* of the funnel. The number of funnels, together with their width, seems to be a much more appropriate measure for characterizing difficult GO problems.

Functions with funnel-shaped landscapes arise in test problems and in important practical problems as well. Many functions from the literature have a single funnel or, at least, the funnel containing the global minimum is a large, easily accessible one; such functions include classical test functions in GO (Rastrigin, Ackley, Levy) and also more recent ones [15]. The same happens in important practical problems arising in chemistry, namely in most — though not all — the Lennard-Jones (LJ) and for some Morse molecular conformation problems — introduced later in the paper — in particular for those instances in which the interaction between particles is "wide-range", and the number of particles is not too large.

There exist in the literature simple but quite effective algorithms which are particularly well suited for functions of the above type: the Basin Hopping (BH) algorithm by Wales and Doye [20] and, the Monotonic Basin Hopping (MBH) algorithm by Leary [9] and some of its variants [1, 2] proved to be extremely efficient in detecting funnel bottoms. However, the solution of GO problems with a large number of funnels and/or a small basin of attraction for the funnel containing the global minimum is a much more difficult task. Such problems include, for example, global minimization with the Schwefel [17] test function and global minimization of hard LJ instances and the Morse instances in the case of short-range interaction.

In this paper we will show that so-called Population Based Approaches (PBAs) are well suited to solve these difficult GO problems. By PBA we refer to a broad class of algorithms which stem from — but are not limited to — the area of Genetic Algorithms (GA). A PBA relies on an evolving collection — *population* — of solutions; evolution is driven by perturbation operators (crossover, mutation) and updating/replacement mechanisms. Such mechanisms embed some device to enforce diversity among members of the population, in order to avoid stagnation.

PBAs are not specifically tailored to GO problems like (1), but offer appealing characteristics:

- they are a powerful tool to perform extensive exploration of the search space, and
- they exhibit remarkable robustness features, even when equipped with unbiased operators, not relying on problem-specific structure.

PBAs have been extensively used for molecular conformation problems – see for example [4] and [6] for the optimization of the LJ potential and [16] for Morse potential. A recent PBA displaying remarkable performance is the Conformational

Space Annealing Method (CSA), proposed in [10]: combining features from GAs and other metaheuristics, this method was able to detect all known putative global optima for the LJ potential for a large range of instances (see [11]).

While MBH-like algorithms are extremely simple, PBAs tend to be more complex and involve an higher amount of technicalities. In this paper we start from the basic template of MBH, minimally extending it to manage a population and incorporating simple but powerful diversity enforcement methods, inspired from the CSA framework, which are based on a *dissimilarity measure*. We call the resulting algorithm PBH (Population Basin Hopping). We formulate several dissimilarity measures and study their impact on the performance of PBH, both on standard test problems and on molecular conformation problems. Computational results show that using a population based approach equipped with a suitable dissimilarity measure can drastically improve performances and robustness over the basic MBH template.

The paper is organized as follows. In Section 2 we give a general description of PBH and discuss some related issues. In Section 3 we compare MBH and PBH over two GO test functions, a single-funnel one, Rastrigin, and the multi-funnel Schwefel function. The results obtained with the Schwefel function are an impressive illustration of how interaction between individuals in the population may lead to a dramatic improvement in the performance of an algorithm. In Section 4 we discuss an important class of GO problems, the molecular conformation ones, by presenting and testing many dissimilarity measures for these problems. In Section 5 we make additional experimental observations. In Section 6 we present an experimental analysis of PBA aiming at a deeper understanding of its behavior. Finally, in Section 7 we draw some conclusions.

2. Basic algorithms

2.1. **Definitions.** We introduce the following definitions. Let $L_f : \mathbb{R}^n \to \mathbb{R}^n$ be a local minimization (descent) procedure returning a local minimum reached from a starting point X (in the following experiments we have always employed the Limited Memory BFGS procedure, see [12]). $\mathcal{S} = \{X = L_f(\bar{X})\colon \bar{X} \in \mathcal{D}\}$ is the set of local minima of f over $\mathcal{D}$; throughout this paper, unless otherwise stated, we will always deal with locally minimized solutions. $Y \in \mathcal{S}$ is defined as adjacent (at level $\bar{\delta}$) to $X \in \mathcal{S}$ if there exists a "small perturbation" $\delta : \|\delta\| \leq \bar{\delta}$ such that $Y = L_f(X + \delta)$. Note that such adjacency relation need not be symmetric; also, it induces a quite natural concept of neighborhood in $\mathcal{S}$.

A sequence of local minima $X_0, X_1, \ldots, X_m \in \mathcal{S}$ such that X_i is adjacent to X_{i-1} $\forall\, i = 1, \ldots, m$ is a descent sequence if

$$f(X_i) \leq f(X_{i-1}) \quad \forall\, i = 1, \ldots, m.$$

A *funnel* corresponds to a maximal set $U \subseteq \mathcal{S}$ such that all maximal descent sequences starting at each $X \in U$ terminate at the same funnel bottom $X^\star \in U$.

We call a *local move* $\mathbf{\Phi}$ a (deterministic or stochastic) procedure that, given a $X \in \mathcal{S}$, generates an adjacent local minimum $Y = \mathbf{\Phi}(X) \in \mathcal{S}$.

2.2. **The MBH algorithm.** The basic structure of MBH, as given in [9] is the following, where MaxNoImprove is a prefixed parameter.

MBH(X **: initial local minimum)**
Step 1. Compute $Y := \mathbf{\Phi}(X)$;
Step 2. **if** $f(Y) < f(X)$ **then** set $X := Y$;
else reject $\bar{Y}$;
Step 3. Repeat Steps 1–2 until MaxNoImprove consecutive rejections have

occurred;
return X;

The local move $\mathbf{\Phi}$ is usually defined as

$$\mathbf{\Phi}(X) = L_f(X + \Delta), \tag{2}$$

where Δ is usually a uniform random vector drawn from a box with given size.

We observe that MBH performs a kind of monotonic depth-first search in $\mathcal{S}$. Despite its simplicity, computational experiments reveal the effectiveness of MBH when faced with GO problems with single funnel landscapes or with a large basin of attraction of the funnel containing the global optimum [9, 1]. In fact, MBH cleverly copes with the structure of a funnel, generating a descent sequence of local minima; the current best solution is heuristically declared to be a funnel bottom after MaxNoImprove non-improving iterations. For this reason we call MBH a *funnel-descent* algorithm.

In order to effectively sample the search space, the algorithm is restarted a number of times from randomly generated starting solutions. This way it behaves like a simple Multistart algorithm applied to a function which corresponds to the objective transformed by means of local searches.

Unfortunately, computational experience shows that effectiveness of MBH lowers for GO problems exhibiting a large number of funnels and/or a global minimum lying at the bottom of a narrow funnel. For these problems, the only way MBH can explore different funnels is by restarting from a new random local minimum once (it believes that) it has reached a funnel bottom, i.e. MBH has to be run in a Multistart fashion. This reveals that MBH is somehow "poor" for what concerns the overall exploration method: the search consists of *pure intensification*, while only quite a crude device — i.e. pure multistart — is left in charge of diversification.

A clever diversification strategy can be crucial for improving performances on harder GO problems. In order to diversify the search without losing the benefits of the funnel-descent features of MBH we can enlarge the scope of search, replacing a depth-first search with a beam-search. Hence we consider to augment the structure of MBH with the simplest instruments taken from population-based approaches: multiple solutions and diversity enforcement.

2.3. **A Population Basin-Hopping Algorithm (PBH).** To define PBH, we need a *dissimilarity* measure $d\colon \mathcal{S} \times \mathcal{S} \to \mathbb{R}^+$ which quantifies the diversity between two solutions. Ideally, $d(X, Y)$ should be close to zero if $X, Y \in \mathcal{S}$ are very "similar": we allow the concept of similarity to be problem-specific; the only essential requirement for d is that $d(X, X) \equiv 0$.

The algorithm goes through the following steps; let K, d_{cut} and MaxStep be prefixed parameters.

PBH

Step 0. **Initialization.** Randomly generate an initial population
$\mathcal{X} = \{X_1, X_2, \ldots, X_K \in \mathcal{S}\}$.

Step 1. **Perturbation.** Compute
$\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_K\}$
by $Y_p = \mathbf{\Phi}(X_p)$, for $p = 1, \ldots, K$.

Step 2. **Sequential replacement.** Repeat for all $Y_p \in \mathcal{Y}$:
let $X_q \in \mathcal{X}$ such that $d(Y_p, X_q)$ is minimum.
if $d(Y_p, X_q) < d_{\text{cut}}$ **and** $f(Y_p) < f(X_q)$ **then**
set $\mathcal{X} := \mathcal{X} \setminus \{X_q\} \cup \{Y_p\}$;
else if $d(Y_p, X_q) \geq d_{\text{cut}}$ **then**

select $X_k \in \mathcal{X}$ such that $f(X_k)$ is maximum, and
if $f(X_k) > f(Y_q)$ **then**
set $\mathcal{X} := \mathcal{X} \setminus \{X_k\} \cup \{Y_q\}$;

Step 3. Repeat from Step 1 for MaxStep iterations, or
until another stopping criterion is satisfied.
return $\langle X \in \mathcal{X}$ having minimum $f(X)\rangle$;

The algorithm is very simple and adds to the MBH template a minimal equipment for managing a population. At each iteration a new set of candidate points (the "children" of the members of the population) is generated (Step 1). Each new candidate Y_p is compared with its closest member X_q in the current population $\mathcal{X}$; Y_p then replaces X_q if their dissimilarity measure is below the threshold d_{cut} and $f(Y_p) < f(X_q)$ (Step 2). If for all elements in $\mathcal{X}$ it holds that $d(Y_p, X_q) \geq d_{\text{cut}}$, than Y_p replaces the worst solution in the current population, provided that it offers a better objective function value; otherwise, Y_p is discarded. Finally, a stopping criterion is checked and if it is not satisfied the main loop is repeated (Step 3).

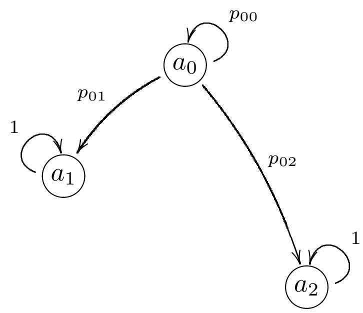
We briefly compare PBH with MBH. What can defeat MBH, if f offers a landscape with many funnels, is that once the search trajectory enters the wrong one — i.e., a funnel not containing the global optimum — it is almost irrecoverably trapped. PBH copes with such phenomenon by maintaining a fixed-size beam of K search trajectories (the evolving population). Such trajectories do not evolve independently: an interaction mechanism, which enforces diversity, is provided at a global level through the similarity-based replacement mechanism (Step 2, see also [11]). We might say that there is *cooperation* between members of the population since the "child" generated by a member may be used to improve a different member of the population. This effect makes PBH drastically different from a set of MBH instances operating in parallel (see also the computational experiments reported in Section 5).

Remarks.

- Note that, independently from d, setting $K = 1$ one obtains the MBH algorithm, hence we see PBH mostly as an extension of MBH.
- The structure of PBH has been kept very simple, — keeping also the same local move of MBH — and quite general. All problem-specific knowledge is concentrated in the definition of $\mathbf{\Phi}$ and d: such operators are open to very special and complex implementations and extensions (e.g. GAs' crossover, directed mutation, etc). In this paper our intention is to keep them as simple as possible and minimize the differences between MBH and PBH in order to let the effect of the population and of its evolution through the dissimilarity measure clearly emerge.

2.4. **Key parameters for Multistart MBH and PBH.** As already pointed out in [5] for Morse problems, the evolution of a run of MBH can be studied from the point of view of Markov chains. Local minima are the states of the chain, while funnel bottoms are its absorbing states. In this section we will consider an extremely simplified Markov chain, which, however, allows us to illustrate the differences between MBH applied in a Multistart fashion and PBH and to emphasize their key parameters. We consider a Markov chain with three states a_0, a_1, a_2 with the following transition matrix:

$$\begin{array}{c|ccc|} & a_0 & a_1 & a_2 \\ \hline a_0 & p_{00} & p_{01} & p_{02} \\ a_1 & 0 & 1 & 0 \\ a_2 & 0 & 0 & 1 \end{array} \tag{3}$$

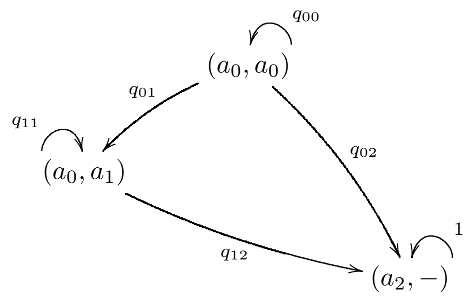


Here states a_1 and a_2 are absorbing states corresponding to two funnels of our problem, and we assume that a_2 is the funnel containing the global minimum. State a_0 is a transient state, which in our problem is the initial state (basically, a random initial point for MBH will very likely belong to a_0 and in the subsequent iterations MBH will evolve towards one of the two absorbing states or funnels). In the Multistart approach we restart MBH until the Markov chain is absorbed into state a_2, thus reaching the global minimum. If we could always stop MBH as soon as an absorbing state has been reached, easy computations show that the expected number of times we need to run MBH before being absorbed in a_2 is equal to $1 + \frac{p_{01}}{p_{02}}$. But one of the key problems in the Multistart approach is to decide after how many iterations we should stop a run of MBH because it is believed that an absorbing state has been reached. Note that in our problem we do not know in advance the absorbing states and we can not simply stop when we reach one of them. We need some rule to (heuristically) decide that we are in an absorbing state. In MBH this rule is given by the MaxNoImprove parameter, i.e. we believe that an absorbing state has been reached when no improvement is observed for MaxNoImprove iterations. Two conflicting objectives are involved when choosing this parameter: computational effort (number of local searches) for a single run of MBH, and rate of success. Of course, the lower the value of MaxNoImprove, the lower is also the computational effort for a single run of MBH. On the other hand, a low value for MaxNoImprove may cause MBH to stop before an absorbing state has been reached thus also decreasing its ability to reach the global minimum. The expected number of times we need to run MBH before reaching a_2 is given by the following function of the MaxNoImprove parameter:

$$\frac{p_{01}+p_{02}}{p_{02}} \frac{1}{1-p_{00}^{\mathrm{MaxNoImprove}}}.$$

Now let us consider a PBH. Even this approach can be studied within the framework of Markov chains. We will consider a population made up by two members but the analysis can be extended to populations of any size. States of the chain are pairs made up by a_0, a_1, a_2. We can reduce the number of these pairs to three: pair (a_0, a_0), i.e. the initial state (similarly to MBH, the initial population will very likely contain two members from a_0); pair (a_0, a_1); pair $(a_2, -)$ (i.e. any pair where one member corresponds to the funnel containing the global minimum). Note that the crucial effect of diversity enforcement is that the chain *cannot* evolve towards the pair (a_1, a_1). The transition probability matrix for this Markov chain is the following

$$\begin{array}{c|ccc|} & (a_0,a_0) & (a_0,a_1) & (a_2,-) \\ \hline (a_0,a_0) & q_{00} & q_{01} & q_{02} \\ (a_0,a_1) & 0 & q_{11} & q_{12} \\ (a_2,-) & 0 & 0 & 1 \end{array} \tag{4}$$



where, taking into account (3) and the effect of the diversification:

$$q_{00} = p_{00}^2 \quad q_{01} = p_{01}^2 + 2p_{00}p_{01} \quad q_{02} = p_{02}^2 + 2p_{00}p_{02} + 2p_{01}p_{02}$$

$$q_{11} = p_{00} + p_{01} \quad q_{12} = p_{02}.$$

Now it can be seen that there is a single absorbing state. In this case the funnel containing the global minimum is within the unique absorbing state. Then, once we have reached the absorbing state, we have also reached the global minimum. As before, also in this case we do not know in advance the absorbing state and we have to decide when to stop the algorithm because we believe that the absorbing state has been reached, i.e. something similar to the MaxNoImprove parameter is needed (this is the MaxStep parameter in PBH). However, there is a fundamental difference with respect to MBH. In MBH if we do not stop a run which has been absorbed into state a_1 not containing the global minimum, we will never be able to reach state a_2 and thus the global minimum. In PBH the absorbing state contains the global minimum and if we let the algorithm run for an infinite number of iterations, we will eventually reach the global minimum. Therefore, while from the practical point of view it may be convenient to stop a run of the population-based approach and restart it from scratch, this is not strictly necessary from the theoretical point of view, while in MBH this is necessary under both points of view. On the other hand, PBH has another key parameter. In our example we have two funnels. For this reason a population composed by two members is enough. However, if the number of funnels increases, then the only way to guarantee that the population-based approach will reach the global minimum is to increase also the size of the population. This quantity should be large enough to guarantee that among the different paths followed by the algorithm at least one will lead to the global minimum. Therefore, population size and diversification tools (the dissimilarity measure d in our case) are key parameters for PBH. We also remark that the best possible choice for the population size is not necessarily equal to the number of funnels in our problem: indeed, as we increase the population size, the expected number of iterations to reach the global minimum is reduced but, on the other hand, the effort per iteration is increased. Which of the two approaches (multistart MBH or PBH) is the most efficient is usually related to the problem at hand. However, our computational experiments will show that in the hardest instances PBH, with a careful selection of the dissimilarity measure, is able to deliver significantly better results than Multistart MBH.

3. Experiments with the Rastrigin and Schwefel test functions

We first discuss the behavior of PBH on some well known test functions in GO, namely the Rastrigin [18] n-dimensional function

$$R_n(x) = \sum_{i=1}^{n} \left[x_i^2 - 10\cos(2\pi x_i)\right] + 10n \qquad x_i \in [-5.12, 5.12],\ i = 1, 2, \ldots, n,$$

and the Schwefel [17] n-dimensional function

$$S_n(x) = \sum_{i=1}^{n} -x_i \sin\left(\sqrt{|x_i|}\right) \qquad x_i \in [-500, 500],\ i = 1, 2, \ldots, n.$$

These functions have a huge (exponentially increasing with n) number of local minima and analytically known global minima. By looking at the 1-dimensional landscapes of the two functions — see Figure 1, it can be observed that, in spite of the huge number of local minima, the Rastrigin function presents a single funnel structure for any n value; on the other hand, the Schwefel function presents relatively fewer local minima (at the same n dimension) but the number of its funnels grows as 2^n. We implemented a PBH algorithm using *minimal* problem-specific knowledge: the $\mathbf{\Phi}$ function is defined accordingly with (2) and is exactly the same employed by MBH, the dissimilarity measure is a simple and quite natural one, the absolute value of the difference between function values:

$$d(X, Y) = |f(X) - f(Y)|, \qquad X, Y \in \mathcal{S}.$$

As we will see soon, this measure is extremely effective for the Schwefel function, but it is important to underline that we are not claiming that it is always the most appropriate one (see also the discussion in Section 4.3.4). We set $d_{\text{cut}} = \infty$, which avoids calibration of the d_{cut} parameter.

The algorithms are compared on the basis of

- NC, the expected number of $\mathbf{\Phi}$ calls before optimum detection – this accounts for the computational effort since $\mathbf{\Phi}$, which requires a local minimization, is the most time-consuming component in the algorithms — and
- RS, the rate of successes (fraction of times the global optimum has been detected) over the total number of runs allowed — this account for robustness and reliability of the algorithms.

Each MBH and PBH run was stopped when the known optimum was found or after a fixed number of search iterations had been performed; this iteration limit was set to 3000 for both algorithms. PBH was run with $K = \frac{1}{2}n$, n, $\frac{3}{2}n$ respectively; a number of 10 runs was allowed for PBH, and 1000 runs for MBH. The results of this series of experiments is reported in Table 1. MBH turns out to be an excellent

	Rastrigin				Schwefel			
	$n = 20$		$n = 30$		$n = 20$		$n = 30$	
	NC_{avg}	RS	NC_{avg}	RS	NC_{avg}	RS	NC_{avg}	RS
MBH	558.1	1000/1000	622.5	1000/1000	–	0/1000	–	0/1000
PBH($K = \frac{1}{2}n$)	5392.0	10/10	9078.0	10/10	14770.0	10/10	219997.5	2/10
PBH($K = n$)	9250.0	10/10	17463.0	10/10	20372.0	10/10	113052.9	7/10
PBH($K = \frac{3}{2}n$)	16323.0	10/10	28804.5	10/10	18441.0	10/10	121516.9	8/10

TABLE 1. Comparison between MBH and PBH.

“funnel-descent” method, and this accounts for its much better performance on the Rastrigin case. In this case the single-funnel landscape basically makes it useless to follow different trajectories and the higher computational effort per iteration of PBH turns out to be a waste. This is confirmed by the experiments, where it is clearly seen that the number of local searches needed for the Rastrigin test functions scales linearly with the number of elements in the population. Much different is the situation for the multi-funnel Schwefel function. When several funnels are present in the searched landscape, an MBH run falling in the wrong one — i.e. one not containing the optimum — has irrecoverably failed. Due to the huge number of

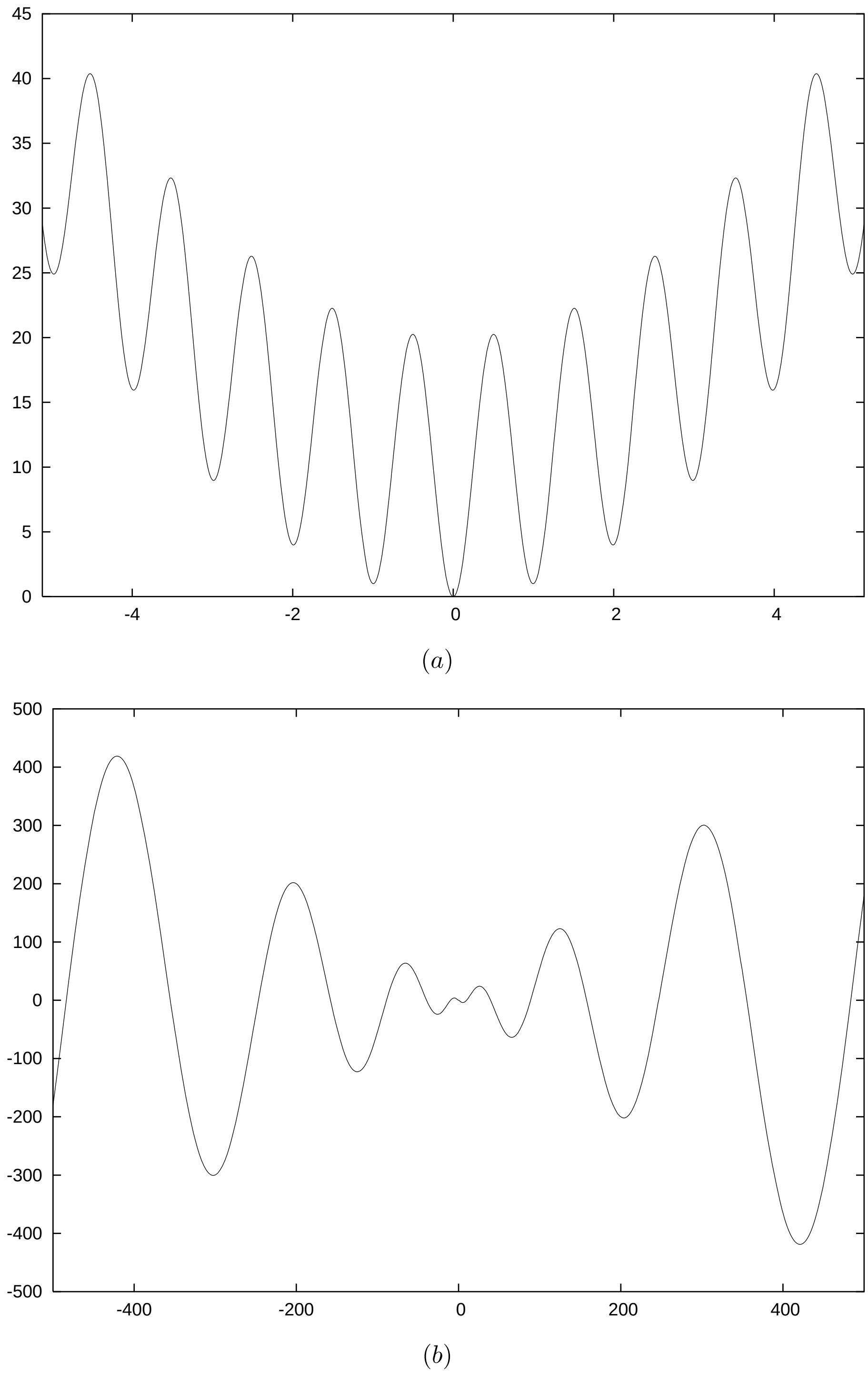


FIGURE 1. 1-dimensional Rastrigin (a) and Schwefel (b) functions.

funnels for the Schwefel function, we need to run MBH a very large number of times (directly proportional to the number of funnels itself) before being able to reach the global minimum (at $n = 20$ MBH was not able to reach the global minimum even after 10,000 runs). On the other hand in PBH diversification enforced by the dissimilarity measure allows to dramatically decrease the effort to detect the global minimum. We note that the number of funnels in the Schwefel case is much larger than the population size, hence there is very little chance for the initial population to have an individual already in the "optimal" funnel. Therefore, only the interaction between the individuals is able to drive a trajectory towards the global minimum.

4. PBH for molecular conformation problems

4.1. **Problem definition.** In this section we will address an important application in the field of global optimization, namely the problem of globally minimizing the Lennard-Jones (LJ) and Morse interatomic potentials, denoted by E_{LJ} and E_{M} respectively. For our purposes, the total interatomic potential of a set of N particles — a "cluster" — is defined by

$$f(X) = E(X) = \sum_{i<j} V(r_{ij}),$$

where $X = (X^1, X^2, \ldots, X^N)$ is a collection of N triplets (x_i, y_i, z_i), each defining the cartesian coordinates of the geometric center of each particle $i \in \{1, 2, \ldots, N\}$, r_{ij} is the Euclidean distance between particles i and j. In the following we will assume that the particles are equal atoms. The *pair potential* $V(r)$ for Lennard-Jones and Morse clusters is given respectively by

$$V_{\text{LJ}}(r) = r^{-12} - 2r^{-6}$$

and

$$V_{\text{M}}(r) = \left[e^{\rho(1-r)} - 1\right]^2 - 1.$$

In the Morse case ρ is a prefixed parameter which allows to model different types of interaction. The problem of determining a minimum energy cluster conformation arises in many different contexts, from chemistry, to material sciences, to biology (see [21] for a recent survey); many properties of molecules depend on their three-dimensional structure, hence it is important to know how atoms aggregate to form complex structures. It is widely believed, and confirmed by experimental observations, that stable structures usually correspond to minimum-energy configurations. E_{LJ} and E_{M} are the simplest, but already significant, energy models for atomic clusters.

Minimizing E_{LJ} or E_{M} over $\mathbb{R}^{3N}$ results in extremely difficult Global Optimization (GO) problems. There is experimental evidence that both potentials exhibit a number of local minima which increases exponentially with N and, even for small-sized clusters, a landscape with a funnel structure is to be searched. While most known putative global optimum configurations of LJ clusters belong to the family of icosahedral based conformations (with a few, significative, exceptions) and can be thought of as being characterized by a single funnel energy landscape, for Morse problems the number of local minima and also the number of funnels increases with the parameter ρ, making the optimization of Morse clusters with large ρ values an extremely difficult task. This has led LJ and Morse problems to become a fundamental testbed for GO techniques.

Our aim is to show that the simple PBH scheme, with an appropriate choice of the dissimilarity measure, is able to improve over a state-of-the-art MBH algorithm

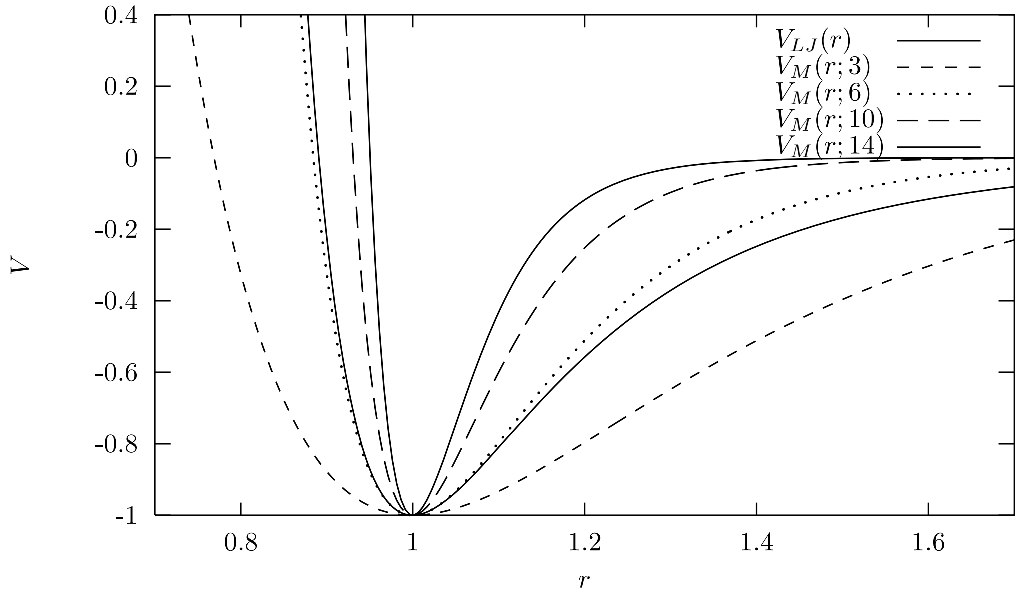


FIGURE 2. Illustration of Lennard-Jones and Morse pair potentials

that reached, to the authors' knowledge, the best performances on very challenging LJ and Morse instances.

4.2. **Two-phase local moves.** Two-phase local moves have been introduced and tested within the framework of MBH both for LJ (see [13]) and for Morse problems (see [5]). A two-phase local move is defined by means of a two-phase local optimization

$$\mathbf{\Phi}(X) = L_E[L_M(X + \Delta)]. \tag{2'}$$

Here M is a function related to E (but hopefully easier to optimize). The inner optimization phase is chosen in order to bias the search towards "promising regions" — i.e. $L_M(X)$ should be ideally very close to the global minimum X^* — while the outer phase projects the search back on the original potential landscape. This amounts to search a reduced set of local minima

$$\mathcal{S}' = \left\{X \colon X = L_E[L_M(\bar{X})],\ \bar{X} \in \mathbb{R}^{3N}\right\} \subseteq \mathcal{S};$$

adjacent minima in $\mathcal{S}'$ are defined accordingly.

From now on we will refer to local minima simply as *clusters*. The experiments presented in [13] for LJ clusters and in [5] for Morse clusters showed that replacing the standard local move (2) with (2′) in MBH allows to detect all the known global optima for LJ instances with $61 \leq N \leq 110$ — in particular, making much easier the solution of the hardest instances, those with non-icosahedral structure — and in the range $41 \leq N \leq 80$ for the extremely challenging Morse instances with $\rho = 14$ (see [5]). To the authors' knowledge, this is up to now the *only* published algorithm which has been able to reach the latter goal.

When employing two-phase local moves, MBH and PBH will be denoted by 2P-MBH and 2P-PBH respectively. For two-phase local moves to be effective, $M(X)$ should be such that the following requirements are satisfied:

R1: $X^* \in \mathcal{S}'$, and

R2: The basin of attraction of X^* is considerably larger in $\mathcal{S}'$ than in $\mathcal{S}$.

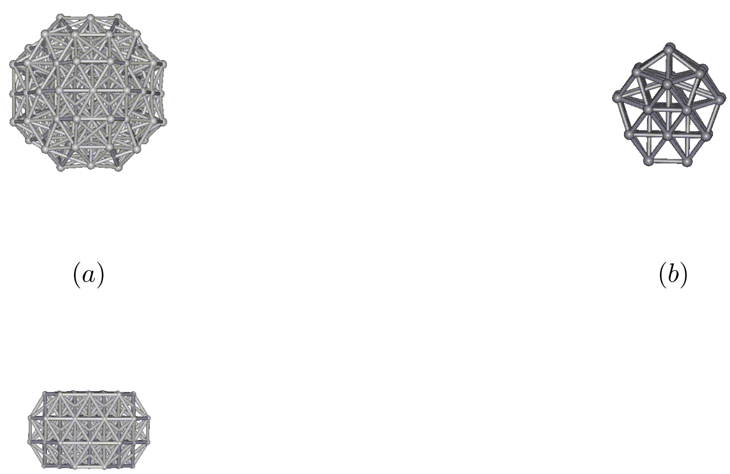
We consider two-phase local minimizations where

$$M(X) = E(X) + G(X)$$

and G is a geometry-based penalty

$$G(X) = \sum_{i<j} \left(\max\left\{0, (x_i - x_j)^2 + w_1(y_i - y_j)^2 + w_2(z_i - z_j)^2 - D^2\right\} \right)^2 + \mu \sum_{i<j} r_{ij}, \tag{5}$$

where $0 \leq w_2 \leq w_1 \leq 1$ and $\mu, D \geq 0$. For a detailed discussion of this kind of penalty and its formulation we refer to [5] and [13]. Here we briefly discuss the choice of parameters w_1, w_2 which are the most relevant ones in the definition of G. Optimal clusters have a more or less regular geometrical shape in the 3D-space. Shapes may considerably differ with the number N of atoms, a fact that is particularly evident for the difficult Morse clusters with $\rho = 14$. Shapes may be spherical, oblate, prolate (see Figure 3). An appropriate choice of the weights w_1, w_2 favors different shapes. For instance, spherical shapes are favored by the choice $w_1 = w_2 = 1.0$, prolate shapes by the choice $w_1 = 1.0, w_2 = 0.7$, oblate shapes by the choice $w_1 = w_2 = 0.65$.



(a) (b)

(c)

FIGURE 3. Cluster shapes: (a) spherical for LJ_{98}, (b) prolate for M_{30}, (c) oblate for M_{61}.

In [5] it has been observed that the best choice for w_1 and w_2 corresponds to the moments of inertia of the optimal cluster X^*. Of course, these quantities are not known in advance and we might choose weights w_1 and w_2 in such a way that one or even both the requirements R1 and R2 are not satisfied. However, 2P-MBH has proven to be robust enough, and a small grid in the parameter space — composed of 2 and 6 points respectively for LJ and Morse clusters — was already containing at least one point satisfying both requirements R1 and R2. As previously remarked, in MBH diversification can only be obtained by random restarting; 2P-MBH is also able to diversify the search by employing different weights in two-phase local optimizations. However, once the weights are fixed, the search in 2P-MBH consists again only of pure intensification as in MBH. Unfortunately, even when weights are chosen that satisfy both the requirements R1 and R2, there still exist suboptimal funnels whose exploration and bottom detection consumes a substantial amount of time. Therefore, the further level of diversification introduced by PBH through the dissimilarity measure turns out to be very useful in many cases.

4.3. **Dissimilarity measures.** Since the core of PBH algorithms is represented by the dissimilarity measure d, we will discuss some possible choices of such measures for molecular conformation problems.

Any appropriate dissimilarity measure should only take into account structural properties of the clusters, allowing to detect and avoid generation of equivalent clusters differing only by rigid transformations like rotations, translations or atom permutations. This, in particular, rules out the standard euclidean distance as a possible dissimilarity measure between clusters. Also it has been experimentally observed that using the absolute difference of the energy of two clusters as a measure of dissimilarity is not significant, as radically different geometrical structures exists with very similar energies.

According to the kind of properties that are taken into account, dissimilarity measures can de distinguished between *local*, which are based on properties of each particle and some of its neighbors, and *global* ones, based on global properties of the atoms.

In what follows we introduce some local and global measures which we tested within the framework of 2P-PBH.

4.3.1. *Measure d_H.* This measure has been adapted from [11]. Consider any atom i of a cluster X, and let its ε-shell be the sphere with radius ε centered at the atom itself: define $H^1(X,n)$ as the number of atoms in X having exactly n other atoms lying in their 1.25-shell; also, define $H^2(X,n)$ as the number of atoms having exactly n between the 1.25- and the 1.55-shell. Then a measure of the distance between two clusters X and Y is defined by

$$d(X,Y) = \sum_{n=0}^{N} n\big[2|H^1(X,n) - H^1(Y,n)| + |H^2(X,n) - H^2(Y,n)|\big]. \tag{6}$$

The parameters 1.25 and 1.55 were empirically determined looking at a statistics on the pairwise distances of Lennard-Jones clusters in the Cambridge Cluster Database. In Figure 4 we report, for different putative optimal clusters, observed pairwise distances: from the figure it is quite evident that distances are clustered around values like 1, which is the minimum of pairwise interactions, and $\sqrt{2}$, which is the diagonal of squares whose edge has unit length. From the figure we can also notice that as soon as, increasing N, the geometry of the cluster radically changes, also the distribution of pairwise distances significantly changes (see for example, the differences between 74 and 75, or 97 and 98), so that an aggregate measure of the difference in pairwise distances can be a good indicator of differences in shape.

This measure is obtained by collecting only information on local properties of the atoms (the number of neighbors of each atom) and is thus a local one.

4.3.2. *Measure d_a.* A global property of an atom is its distance from the center of mass of the cluster. This information can be collected into the following index for

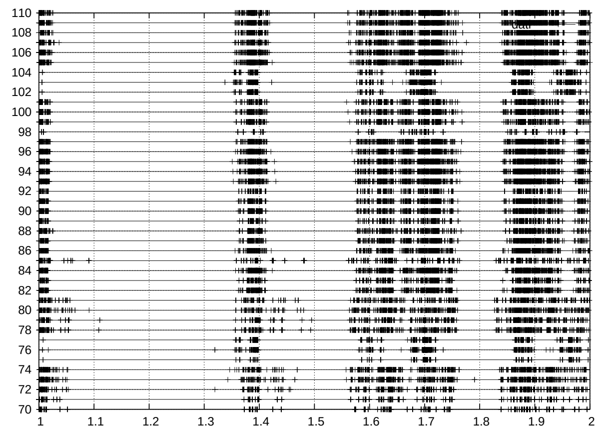


FIGURE 4. Pairwise distances in Lennard-Jones clusters

a given cluster X:

$$I_a(X) = \sum_{i=1}^{N} \exp(a\|X^i - X^0\|),$$

where X^0 is the cluster's center of mass. The value a might be either positive or negative. If a is negative, then atoms in the cluster are given a weight which decreases as their distance from the center of mass increases, while for a positive a value, the weight increases with the distance from the center of mass. The magnitude of a controls the rate at which the weight decreases or increases. Now we can define a whole class of global dissimilarity measures indexed by a. Given two clusters X and Y, measure d_a is simply the absolute value of the differences between the corresponding indices:

$$d_a(X,Y) = |I_a(X) - I_a(Y)|. \tag{7}$$

If a is negative, measure d_a mainly depends on dissimilarities between the kernels of the clusters, while for positive a values it mainly depends on dissimilarities between the external layers of the clusters.

4.3.3. *Measure* d^p_{ord}. It is important to remark that the same information can be exploited in several ways to define a dissimilarity measure. The global measure d^p_{ord} is based on the same information as d_a, namely distances of atoms from the cluster's center of mass, but this information is employed in a different way. First, for a given cluster X, the distances are sorted in a nondecreasing way and stored in a vector ord_X. Note that this way we are adding further information for each atom, since we do not simply store its distance from the center of mass but also the ranking of this distance with respect to those of the other atoms. Then, given two clusters X and Y, we define the following class of dissimilarity measures indexed by the positive value p:

$$d^p_{ord}(X,Y) = \sum_{i=1}^{N} |ord_X[i] - ord_Y[i]|^p .$$

Note that parameter p controls the impact on the whole measure of the differences between distances with the same ranking in the two clusters. Indeed, the relative impact of a larger difference with respect to a smaller one increases as p increases.

4.3.4. *Other measures.* In measures d_a and d_{ord}^p we have collected information about distances of single atoms with respect to the center of mass of the cluster. Another possible information is represented by the set of all distances between atom pairs. Though not explored in this paper, measures d_a and d_{ord}^p could be easily redefined by substituting distances from the center of mass with distances between all atom pairs. Actually, the simple dissimilarity measure employed in Section 3, namely the difference between the energy values of two clusters (also already employed in [4] for LJ problems), is based on all the distances between atom pairs. The energy value of a cluster can be seen as an index associated to a cluster. Such index depends on all distances between atom pairs, which are given a weight equal to the energy contribution of the pair. Then, a higher weight is given to distances close to 1, which makes the corresponding measure more of a local nature than of a global one. In our computational experiments we tested this distance but, as expected, we did not get good results. Indeed, while for the Schwefel function (on which this dissimilarity measure works extremely well as reported in Section 3) closeness of the function value at a given local minimum to the global minimum value also corresponds to closeness of the local minimum to the global minimum point in the search space, when Morse clusters with large ρ value are considered, closeness in energy values is often not related to closeness between structures, so that in many cases clusters are erroneously considered as similar by this measure.

Hartke [6] defines a cluster "shape index" g whose computation is based on the two-dimensional projection of the cluster (see reference for details). This value turns out to be quite different for geometrically different cluster structures (FCC, decahedra, icosahedra) and this allows to clearly distinguish the geometrical structure of the cluster. Then, a possible distance is $d_g(X, Y) = |g_X - g_Y|$. However, according to our experiments this measure does not seem to be an efficient one for Morse clusters with large ρ value. This is due to the fact that in these Morse clusters the search is often driven towards suboptimal clusters which have the same geometrical structure as the optimal one and, consequently, the measure based on g values is not able to distinguish between them.

4.4. **Computational results.** In this section we present the computational results for 2P-PBH with different dissimilarity measures and compare it with those of other algorithms, in particular 2P-MBH.

4.4.1. *Selected test cases.* In order to assert the impact of the dissimilarity measures within the PBH framework, we first tested the algorithms over a small, but carefully selected, set of six difficult instances. In particular, we have considered: a single, very challenging, Lennard-Jones cluster, the one with $N = 98$ atoms (LJ_{98}), whose new putative global minimum has been only recently detected and has a very special geometrical shape [8, 9]; five Morse instances at $\rho = 14$, namely M_{30}, M_{43}, M_{55}, M_{61}, M_{79}, which are representatives of classes for which a very low percentage of successes was observed in [5] for 2P-MBH. After this comparison over difficult instances, more extensive computations on Morse instances have been performed only with the dissimilarity measure which has turned out to be the most robust one (see Section 4.4.4).

4.4.2. *Tested algorithms.* We have tested 2P-PBH with different dissimilarity measures and compared with the results of 2P-MBH and, when available, MBH and the already mentioned population based approach CSA (see [11]). We recall that any algorithm employing two-phase local searches needs parameters for the two-phase optimization to be specified; we used on each tested instance the best parameter set identified in [5, 13] and reported in Table 2(a). As also previously remarked, this is not a severe limitation since a small grid in the parameter space is sufficient to handle most instances; the computational effort devoted to gridding is usually largely compensated by the quick convergence of the algorithm. However, we always have to recall that the number of local searches for algorithms employing two-phase local searches should be multiplied by the number of different parameter sets employed (only 2 for LJ, 6 for Morse). For what concerns 2P-MBH we have set the MaxNoImprove parameter to 1000 for LJ_{98}, to 200 for $M_{30,43}$, and to 500 for $M_{55,61,79}$. In 2P-PBH the maximum allowed number of iterations MaxStep was set equal to 1500. However, we remark that this limit only came into play for M_{79}, in all the other cases the global minimum was always detected well before reaching this limit. In 2P-PBH the population size K is set to 40 for $\mathrm{LJ}_{98}, M_{30,43}$, 80 for $M_{55,61}$ and 160 for M_{79}. The larger population for Morse instances with respect to LJ ones (also taking into account the lower number of atoms) is justified by the fact that the energy landscape for E_{M} at $\rho = 14$ is rougher than the one for E_{LJ} and, consequently, more monotonic paths should be followed in order to include one leading to the global minimum. Both for 2P-MBH and 2P-PBH the perturbation Δ is a random one into the box $[-0.4, 0.4]^{3N}$. Within the 2P-PBH framework we tested the following dissimilarity measures: d_H, d_a with $a = \pm 1$ (no other a value has been tested), d^3_{ord} (also the neighbor values $p = 2, 4$ have been tested and the overall results were only slightly inferior). Accordingly with [11], the d_{cut} values have been fixed, after a quite limited parameter tuning carried on the smallest Morse instances, proportionally to the average distance d_{ave} in the initial population: $1.5d_{\mathrm{ave}}$, $0.1d_{\mathrm{ave}}$, $0.05d_{\mathrm{ave}}$ and $0.5d_{\mathrm{ave}}$ respectively for d_H, d_{-1}, d_{+1} and d^3_{ord}.

4.4.3. *Computational experiments and their interpretation.* Table 2(b) summarizes the results obtained on selected test instances. Aside from the table, we mention the performances of other effective algorithms applied to molecular conformation problems. The original MBH as defined by Leary [9] relies on the basic local move (2). On LJ_{98}, MBH detects the putative global optimum in 6 out of 1000 runs with an expected number of local searches equal to 180,000, while CSA detects it in 10 out of 10 runs with an expected number of local searches equal to 328,240. We recall that the results for methods including two-phase local searches should be multiplied by the number of parameter sets employed (these were only 2 for LJ clusters). Even taking into account multiple runs, the results confirm what was already observed in previous works, i.e. that two-phase local searches allow a strong improvement in the efficiency when dealing with hard LJ instances. About Morse instances, algorithm CSA has not been applied to them, while MBH has been but often a very large number of runs (up to 10,000) were required to first observe the global minimum for Morse clusters at $\rho = 14$ with up to $N = 80$ atoms, and in some cases the global minimum was not even reached. Therefore, for the very challenging Morse clusters at $\rho = 14$, two-phase local searches appear to be essential for the success of the approach.

In most cases, using 2P-PBH results in a reduced $\mathrm{NC}_{\mathrm{avg}}$ and a higher RS with respect to 2P-MBH; this accounts for increased efficiency and reliability in the PBH framework. On the other hand, the dissimilarity measure seems to play a crucial role for achieving best performances, especially on the hardest instances. This is especially true for the difficult M_{79} test where, with some measures, 2P-MBH still outperforms 2P-PBH.

Effectiveness of a dissimilarity measure is a very problem-specific issue and, for this problem, it seems also related to each instance — specifically, to the geometry of the optimum. The following discussion is tailored to molecular conformation problems, but can be extended to more general contexts. Ideally, a dissimilarity measure should always be able to clearly distinguish between the optimal cluster (and any cluster from which the optimal one is easily accessible) and all the suboptimal clusters from which the optimal one is not accessible. Given the extreme variability of the optimal geometries as the number N of atoms changes, it is unlikely that an “optimal” dissimilarity measure, whose efficiency is always superior to all the others, exists; this would require to identify a *unique* global or local property with respect to which clusters should be compared. On the other hand, it is possible to search for *robust* measures, i.e. measures which always guarantee a good performance (although not necessarily the best one in all cases). Our computations seem to indicate that measure d^3_{ord} is a quite robust one.

The d_H measure seems to be well suited for the LJ problem, and also gives good results on M_{30}–M_{61}, but the heavy failures on M_{79} suggest a lack of robustness. With d_H we have also tested a version with annealing of the d_{cut} value (use of annealing is envisaged in [11] for CSA): d_{cut} is multiplied by 0.85 every 50 iterations during the first 250 iterations. Interestingly enough, when the d_{cut} value is subjected to annealing an improvement on some — not all — instances is observed. This may be explained as follows: together with the dissimilarity measure, the d_{cut} parameter allows to distinguish different geometries, hence influencing the ability of the algorithm to resolve the optimum cluster among suboptimal ones. Annealing may lead to modify the d_{cut} to more appropriate values during the same run; anyway setting the best possible d_{cut} value may be as difficult as selecting the optimal dissimilarity measure. The same annealing procedure tested with the other measures did not deliver satisfactory results, hence we pursued experiments with a fixed, previously calibrated d_{cut}.

The overall results for d_{+1} and, even more, for d_{-1} are not very satisfactory. However, we decided to include these measures and to report their results because in some cases their behavior is clearly superior with respect to other measures. This agrees with the above discussion, and confirms that a measure based on the “right” property really allows a remarkable improvement of the efficiency. Unfortunately, the “right” property may be considerably different at different sizes and this is basically what makes impossible the search for an optimal measure.

According to the computational results presented above, measure d^3_{ord} appears to be the most robust one. It is important to investigate the reasons of its success with respect to other measures. In what follows we propose some conjectures to explain this.

Measure $d_{\pm 1}$ are biased dissimilarity measures, in the sense that dissimilarity between two clusters is not uniformly measured but a higher weight is given to dissimilarity in some parts of the clusters with respect to others. In particular, as already remarked in Section 4.3.2, d_{-1} mainly depends on dissimilarities between the kernels of the clusters, while d_{+1} mainly depends on dissimilarities between the external layers. Consequently, these measures may work very well (even much better than other measures) when the difference between the optimal and suboptimal clusters mainly lies in the parts of these clusters on which they depend, but may perform quite badly if the difference mainly lies in other parts. This is indeed the kind of behavior, previously commented, of measures $d_{\pm 1}$. Measure d^3_{ord}, although based on the same information as $d_{\pm 1}$ (distances of atoms from the center of mass), is not a biased one in the sense given above. Differences at each layer of the clusters are weighted in the same way. This unbiasedness is probably the reason for the higher robustness of this measure with respect to $d_{\pm 1}$.
A comparison between measures d_H and d^3_{ord} is more difficult because these are essentially different measures. Indeed, as already observed in Section 4.3, d_H is a local measure, while d^3_{ord} is a global one. Our experiments suggest that, especially

Test	μ	D	w_1	w_2
LJ_{98}	0.0	3.5	1.00	1.00
M_{30}	0.0	2.5	1.00	0.70
M_{43}	0.0	2.5	0.75	0.50
M_{55}	0.1	3.5	1.00	1.00
M_{61}	0.0	2.5	0.65	0.65
M_{79}	0.1	4.0	1.00	1.00

(a) Parameter settings for each instance.

Test	LJ_{98}		M_{30}		M_{43}		M_{55}		M_{61}		M_{79}	
	NC_{avg}	RS	NC_{avg}	RS	NC_{avg}	RS	NC_{avg}	RS	NC_{avg}	RS	NC_{avg}	RS
2P-MBH	11 347	7/50	13 648	12/500	11 425	16/500	41 987	10/500	53 787	8/500	110 627	5/500
2P-PBH(d_H)	8 440	10/10	6 468	10/10	2 992	10/10	15 392	10/10	28 528	10/10	534 080	4/10
2P-PBH*(d_H)	4 197	10/10	5 420	10/10	5 216	10/10	15 496	10/10	40 092	10/10	345 440	5/10
2P-PBH(d_{+1})	10 200	10/10	5 400	10/10	12 416	10/10	8 084	10/10	16 880	10/10	320 745	4/10
2P-PBH(d_{-1})	20 128	10/10	10 224	10/10	2 988	10/10	7 000	10/10	19 000	10/10	-	0/10
2P-PBH(d^3_{ord})	8 720	10/10	5 148	10/10	7 496	10/10	16 984	10/10	17 296	10/10	48 228	10/10

(b) Computational results. PBH* utilizes annealing on d_{cut}.

TABLE 2. Computational results on selected instances.

for Morse clusters, a measure which only depend on local properties might not be appropriate in some cases. In particular, for M_{79} we computed the dissimilarity between the optimal cluster and other suboptimal clusters measured by d_H and this was quite low, which means that this measure does not seem to be able to clearly distinguish between the optimal and suboptimal clusters.
We underline once again that the above explanations are tentative and not rigorous ones and any further investigation would be more than welcome not only to understand the behavior of the proposed measures but also to possibly develop other, even more robust, ones.

4.4.4. *Extensive computations with the d^3_{ord} measure.* Once we have experimentally established that measure d^3_{ord} is the most robust one over the previously tested hard instances, we further tested the robustness of this measure by running more extensive tests with it. In particular we run 2P-PBH with all the six parameter sets employed in [5] and reported in Table 3(a) over all the very challenging Morse instances with $41 \leq N \leq 80$ at $\rho = 14$. We set $K = 40$ up to $N = 50$, $K = 80$ from $N = 51$ up to $N = 70$, and $K = 160$ from $N = 71$ up to $N = 80$. In Table 3(b) a $*$ denotes the ability of 2P-PBH of reaching the global minimum for the given number N of atoms and we notice that all global minima in this range have been reached with at least one (and often more than one) parameter set. This further confirms the robustness of the measure.

4.5. **New putative optima.** Finding new putative optima for difficult global optimization like those related to molecular conformations cannot be considered as a priority in judging the goodness of an algorithm – trivially because, even if no proof of optimality exists, global optima might have been already discovered. However when, as in our case, not only an algorithm detects all previously known optima, but it is capable of finding new configurations whose energy is lower than the best known so far, this appears to be a further confirm of the capabilities of the method. For what concerns the algorithm presented in this paper, already in its first "sequential" version (i.e., not based on a population) presented in [5] it was possible to discover 5 new putative optima at $\rho = 8$ and one at $\rho = 6$. While executing the experiments reported in this paper, further 4 new putative optima were detected for very high values of ρ. It has to be remarked that these new conformations were obtained while looking for optimum configurations at $\rho = 14$. Indeed, it is always a reasonable practice to check whether the configurations obtained at the end of a computational experiment are optimal for different values of ρ. This trial is quite cheap, as it amounts just to occasionally run a single local optimization starting from each of the elements in the current population, for different values of ρ. The intrinsic characteristics of population-based methods in which diversity among members is enforced tends to produce quite different geometrical structures that, in some cases, although not optimal for the particular value of ρ used in the experiments, are very close to good optima for different ρ values. This is indeed the case quite often, and we observed that often among the elements of the final population we could find the putative optima for "standard" ρ values (3,6,10). For each of the structures deposited, the Cambridge Cluster Database [3] indicates a range of ρ values for which the structure is supposed to be optimal. We checked the structures obtained during our experiments and find 4 new putative optima. In the following table we report the number of atoms of the new putative optima as well as the range of values of the ρ parameter for which their energy appears to be

	D	μ	w_1	w_2
$P1$	3.5($N \leq 60$)-4	0.1	1	1
$P2$	3($N \leq 60$)-3.5	0	1	0.7
$P3$	2.5($N \leq 60$)-3	0	0.8	0.8
$P4$	2($N \leq 60$)-2.5	0	0.75	0.5
$P5$	2($N \leq 60$)-2.5	0	0.65	0.65
$P6$	2($N \leq 60$)-2.5	0	0.5	0.5

(a) The six sets of parameters employed in two-phase local searches.
The dependence of the parameter D on cluster size is given in parentheses.

N	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
$P1$											*	*		*	*	*	*	*	*	*			*	*				*	*	*	*	*	*	*	*	*			*	*
$P2$									*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*		
$P3$										*		*		*	*	*		*	*	*				*					*	*	*	*	*	*	*	*	*			
$P4$	*	*	*					*	*	*	*	*											*	*	*				*	*	*	*	*	*	*	*	*	*		
$P5$								*	*	*	*	*	*					*			*								*	*	*	*	*	*	*	*	*	*		
$P6$			*	*	*	*	*	*	*																									*	*	*	*	*		

(b) A $*$ denotes a success of 2P-PBH in reaching the global minimum with the given parameter set.

TABLE 3. Extensive computational experiments.

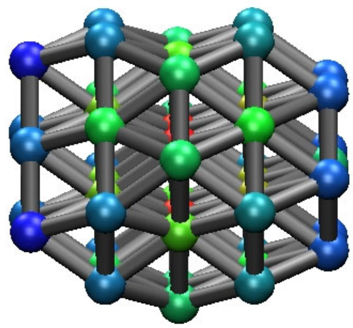


FIGURE 5. Morse 48E

strictly lower than that reported in the database:

N	ρ
48	≥ 20.22
64	≥ 24.31
67	≥ 21.85
72	$\in [19.13, 21.49]$

In Figures 5–8 the geometry of these new optima are displayed. In order to stress the fact that the new putative optima are indeed different from those deposited in the CCDB, we report in the following table a comparison between the geometries of our proposed optima and those found in the database. Under the headings Vert, Surf, Int, we reported, respectively, the number of atoms which are vertices of the convex hull of the cluster, the number of atoms which are on the surface, without being vertices, and the number of internal atoms. The differences clearly indicate that, although the energy of the new putative optima is only slightly lower than that reported in the database, the geometry, at least for some of the new structures, changed in a quite significative way.

Structure	Vert	Surf	Int
$48D$	21	18	9
$48E$	28	8	12
$64D$	28	18	18
$64E$	30	15	19
$67F$	27	24	16
$67G$	27	22	18
$72E$	29	20	23
$72G$	25	25	22

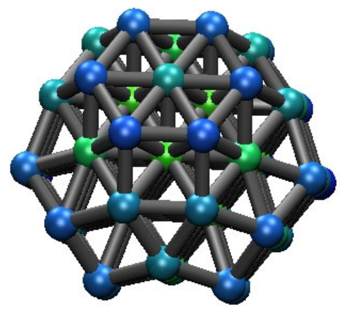


FIGURE 6. Morse 64E

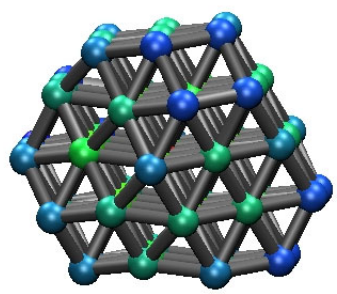


FIGURE 7. Morse 67G

5. Further observations

Some care is needed when comparing different approaches. As we underlined in Section 2, a key parameter for the Multistart approach 2P-MBH is MaxNoImprove. Therefore, one possible question is the following: are our new and better results simply due to a bad choice of the MaxNoImprove parameter in 2P-MBH? Basically, we are asking ourselves when it is convenient to stop and restart the algorithm (a question which can be extended to other GO algorithms, see e.g. [7]). The

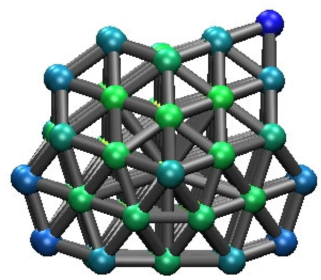
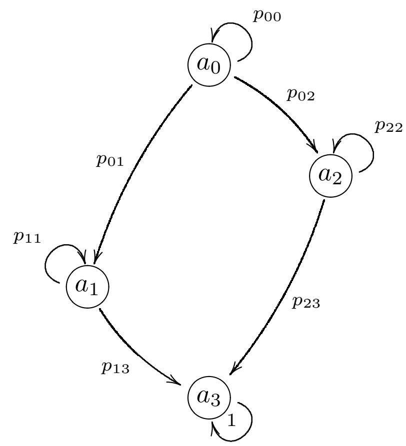


FIGURE 8. Morse 72G

results in Table 2(b) for LJ_{98} have been obtained with MaxNoImprove = 1000, the same choice as in [9], but it is possible that other values could deliver much better results. In order to test this, we started further 100 runs of 2P-MBH collecting the results for different values of MaxNoImprove ranging from MaxNoImprove = 100 to MaxNoImprove = 4000. As expected, the number of successes increased with the value of MaxNoImprove, moving from 2 successes for MaxNoImprove = 100 until the 50 successes for MaxNoImprove = 4000. The best average number of local searches to first hit the global minimum was approximately 9400 local searches, which is still worse than the results obtained with the population-based approaches based on the dissimilarity measures d_H (with and without annealing) and d^3_{ord}. But it is also interesting to observe the following fact. The best average results were obtained with two rather different and extreme values for MaxNoImprove, namely MaxNoImprove = 100 and MaxNoImprove = 3600. Basically, the best strategies are somehow opposite ones, one strategy, the one with MaxNoImprove = 100, made up by very short runs, the other strategy, the one with MaxNoImprove = 3600, with very long runs. This fact can be explained as follows. For LJ_{98} with the given set of parameters the situation is different from the one described by the Markov chain (3) and is better described by means of the following Markov chain:

$$
\begin{array}{c|cccc|}
 & a_0 & a_1 & a_2 & a_3 \\
\hline
a_0 & p_{00} & p_{01} & p_{02} & 0 \\
a_1 & 0 & p_{11} & 0 & p_{13} \\
a_2 & 0 & 0 & p_{22} & p_{23} \\
a_3 & 0 & 0 & 0 & 1
\end{array}
\tag{8}
$$



In this case, the energy landscape, once transformed by means of the two-phase local searches, is characterized by a single funnel/absorbing state, namely state a_3; this funnel, however, can be reached through different paths, i.e. $a_0 \to a_1 \to a_3$ and $a_0 \to a_2 \to a_3$. In the nasty cases we have $p_{01} \gg p_{02}$, i.e. starting from state a_0 it is much easier to evolve towards state a_1 than towards state a_2, but, on the other hand, $p_{13} \ll p_{23}$, i.e. it is much more difficult to evolve towards the global minimum when we are in state a_1 than when we are in state a_2. This is exactly what happens with LJ_{98} where state a_1 is represented by a cluster at an energy level equal to -543.546725 (the global minimum is at the energy level -543.665361) from which it is very difficult to evolve towards the global minimum. This also means that the best value for MaxNoImprove depends on the path followed. This justifies the fact that there are two best choices for MaxNoImprove, i.e. MaxNoImprove = 100 and MaxNoImprove = 3600. While the former value is the most appropriate if we follow easy paths towards the global minimum ($a_0 \to a_2 \to a_3$ in the example), the latter one is the most appropriate if we follow difficult paths ($a_0 \to a_1 \to a_3$ in the example). The advantage of 2P-PBH is that the diversification allows to follow many *distinct* paths towards the global minimum, including the easy ones. This way 2P-PBH usually reaches the global minimum in a limited number of iterations, thus compensating the larger effort per iteration. The situation described above is quite different with respect to the one of the Rastrigin function described in Section 3, even if also the latter has a single funnel. Due to its high symmetries, in the Rastrigin function distinct paths to the global minimum are of comparable difficulty and in this case the larger effort per iteration of PBH is not compensated by a strong reduction of the number of iterations to reach the global minimum.

Another possible doubt about 2P-PBH is that what really matters is only the parallelism, i.e. following many paths in parallel, while the diversification, enforced through the dissimilarity measure, has no impact on the good performance of the population-based approaches. In order to test this we run a population-based approach with the same population size for the different number of atoms but without any diversification (basically, a number of independent runs of 2P-MBH equal to the population size are run in parallel). The results obtained with this approach show that it is less robust than 2P-PBH (e.g. for M_{61} in 2 out of 10 cases it was not able to reach the global minimum within 1500 iterations) and it is even less efficient than 2P-MBH. Similar results were obtained by repeating the experiment with the Schwefel test function. This further confirms the importance of following *diversified* paths in the population-based approaches.

6. An experimental analysis of PBH

We now discuss experimental observations made on PBH, whose goal is to point out relevant characteristics of the search process, in order to justify the improved performance.

First of all we stress that enforcing diversification by way of a dissimilarity measure is *fundamental;* as already remarked in Section 5, modifying Step 2 so that each Y_p is compared only with the X_p from which it has been generated sensibly worsens performance (see also the final paragraph in Section 5).

By tracking the search and investigating the trajectories, we identify two kinds of phenomena.

Survival. This happens when $Y^p = \Phi(X_p)$ with $f(Y_p) < f(X_p)$ but, because of higher similarity, Y_q is compared with $X_q \neq X_p$ and does not replace X_p. This way X_p — which would be overwritten in a monotonic, single-trajectory search — has a chance to *survive* in the next generation, while a different trajectory stems from Y_p if it replaces X_q.

Backtrack (uphill) moves. We call a backtrack move the process by which $f(Y_p) \geq f(X_p)$, but $f(Y_p) < f(X_q)$ and Y_p replaces X_q because of similarity. Remarkably, this way non-improving steps can be incorporated into the search, although the population as a whole *monotonically* evolves — suppose X_p and X'_p are solutions in position p in the population array, with X'_p belonging to an iteration successive to that of X_p: then we will always have $f(X_p) \geq f(X'_p)$.

In order to better appreciate these effects, we track the paths driving to the putative global minimum for some runs of the population-based approach over distinct test cases. We track the optimal trajectory $f(X^*)$ (labelled OPT) and *secondary* trajectories that stem from the optimal one because of survival and backtracking events. Observations on Morse problems are discussed first. Figure 9 points to the role of the survival phenomenon. Survival happens at iterations 2, 6, 36, 37. In all cases $f(X^*) > f(Y^*)$, but X^* is not deleted because Y^* is driven to replace a different member; this bifurcates the search originating the secondary trajectories A, B, C, D. Note that all secondary trajectories *quickly* improve, then suddenly get stuck at non-optimal points; *note:* trajectory B gets stuck at -106.7653718 which is not the optimum (located at -106.8357897).

Figure 10 shows how backtrack moves work; backtracking happens at iterations 4, 252, 449 (note the strongly uphill move at step 252) where $f(X^*) < f(Y^*)$ but, due to similarity, Y^* replaces another member. The optimal trajectory incorporates the backtrack move, while the secondary ones get stuck again: trajectories C, D and E track the non-backtracked solutions. Other bifurcations are due to survival, like in Figure 9.

Apparently, for Morse problems, backtracking happens less often than survival, but still it plays a crucial role in several significant examples.

A somehow different behavior is observed by tracking search paths for runs on the Schwefel function; Figure 11 shows a run for the $n = 20$ case. For this problem, the most frequent phenomenon is backtracking; this is reasonable, since an effective algorithm must have the ability to jump across the many funnels present in the landscape. The survival phenomenon is still present, and again plays a crucial role; the tracked trajectory for example presents 16 backtrackings, and 9 survivals that prevent the optimal trajectory from being erased by one of its own “children” with a better objective function but not leading to the optimum. Secondary trajectories are not shown in Figure 11, for conciseness.

We can conclude that both survival and backtracking are essential for the success of PBH because they prevent the search from converging too quickly to points which

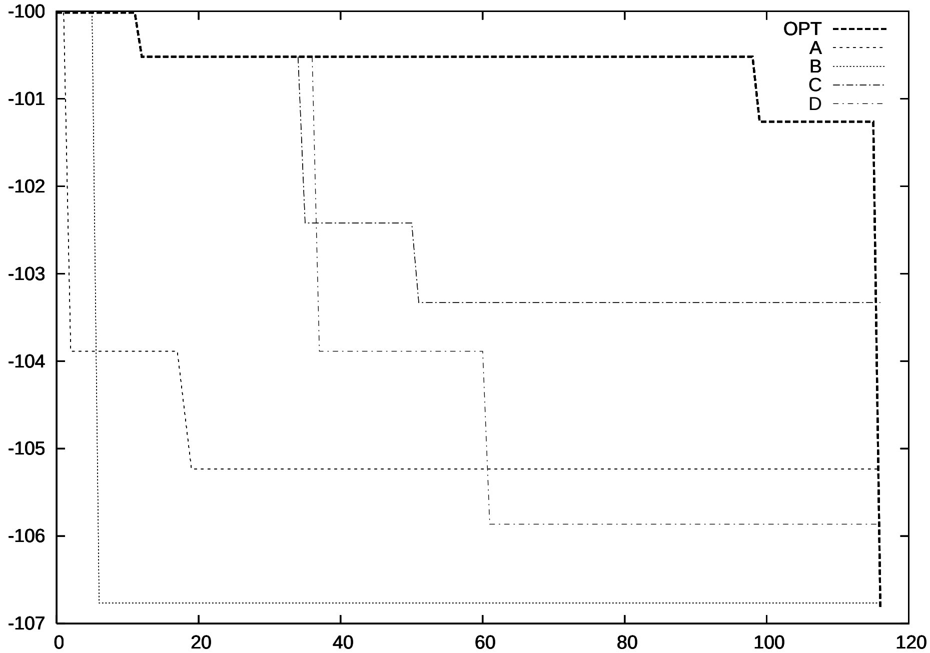


FIGURE 9. Search trajectories, PBH(d_H) on M_{30}.

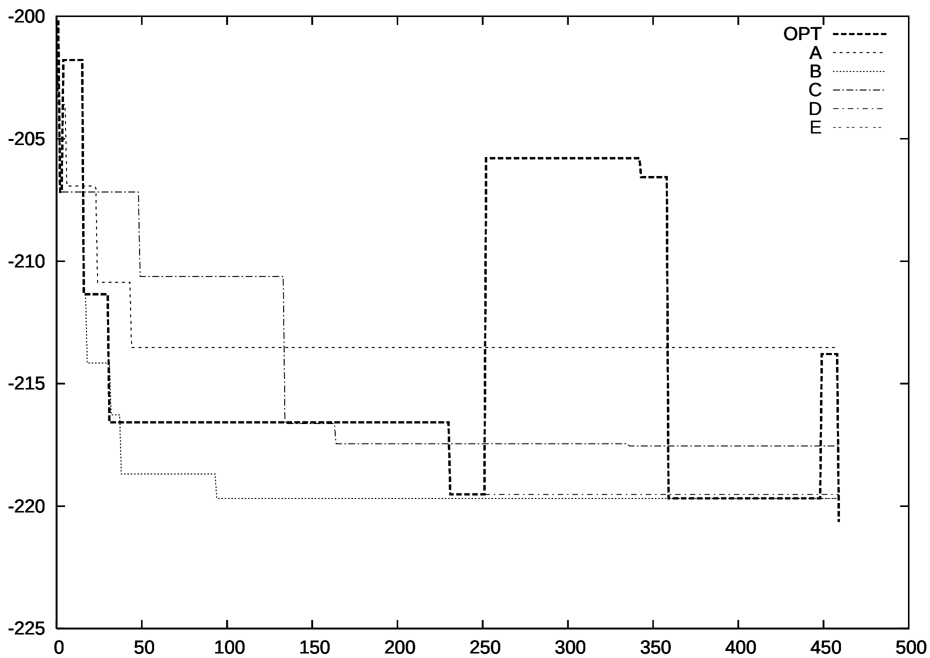


FIGURE 10. Search trajectories, PBH(d^3_{ord}) on M_{55}.

are almost optimal (i.e. whose function value is very close to the optimal one) but from which we are not able to reach the global minimum (particularly, non-optimal funnel bottoms).

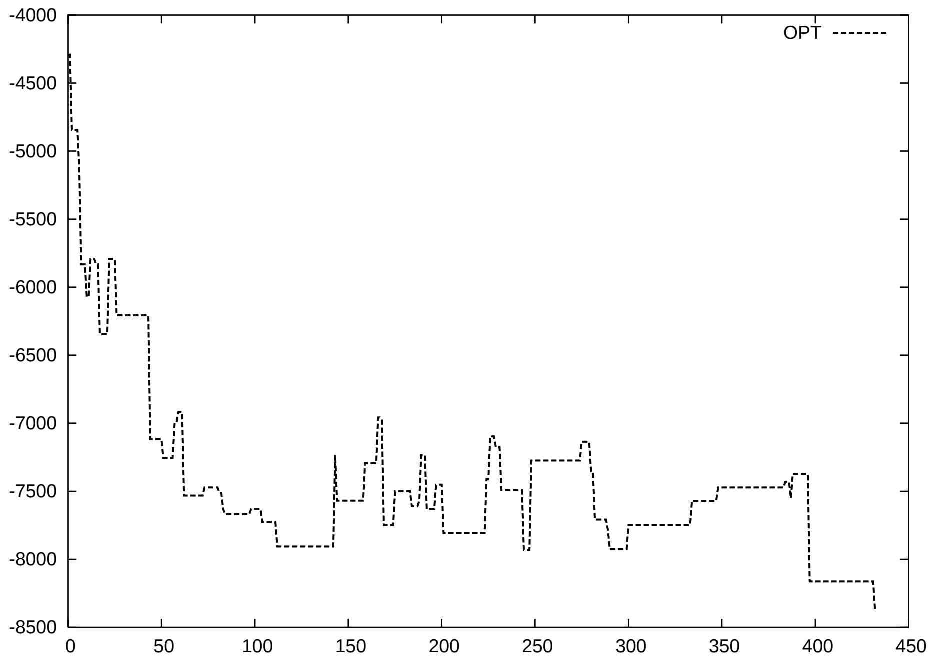


FIGURE 11. Search trajectories, Schwefel functions with $n = 20$ variables.

7. CONCLUSIONS

We have presented a detailed analysis on the significant performance of quite standard global optimization algorithms cast into a population-based framework. Our results, obtained on well known extremely hard large scale global optimization problems, show that if a suitable dissimilarity criterion is used, an impressive enhancement of both speed and robustness can be achieved. Among the examples presented in this paper we would like to stress the importance of the capability of finding all known putative optima for Morse clusters at $\rho = 14$ up to 80 particles, a result which no other single algorithm published up to now has been able to reach, as it is stated in the official repository of putative optima cluster configuration (the Cambridge Cluster Database [3]) where it is affirmed that "Morse clusters provide [...] a rigorous test of a global optimization algorithm. The location of all the global minimum at $\rho = 3$, 6, 10 and 14 for clusters with up to 80 atoms would be a significant achievement and one which no unbiased global optimization algorithm has yet managed". In this paper we presented the results obtained for $\rho = 14$ and $N \in [41, 80]$ only for sake of conciseness: the remaining cases, either for smaller ρ or for smaller number of atoms, are considerably easier to solve – this task can be quite easily accomplished with the algorithm presented in this paper.

A possible limitation of PBH is its lack of generality: it is difficult (probably it is impossible) to find a dissimilarity measure which works well for all GO problems. This was clearly observed and commented when discussing dissimilarity measures for molecular conformation problems and is also confirmed by the fact that most of these measures are strictly problem-specific and the only general one, namely the difference between function values, works extremely well with the Schwefel function but did not give good results on the hard Morse instances. However, in our view this is not a real limitation but a mere consequence of the great variety of GO problems, which prevents from having a *unique* measure working well for *all* GO problems. On the other hand, our computational results on hard GO problems seem to indicate that, though nontrivial, the definition of a good dissimilarity measure for a given class of GO problems may be extremely rewarding in terms of computation times.

Acknowledgments

The research of M. Locatelli and F. Schoen was partially supported by Progetto FIRB "Ottimizzazione Non Lineare su Larga Scala". The authors acknowledge also the contribution of V. Crini, Università di Firenze, for his support in the development of the software, of B. Addis, for many fruitful discussions, and G. Obertelli, University of California, Santa Barbara, for making computing facilities available.

References

[1] B.Addis, M.Locatelli, F.Schoen, Local optima smoothing for global optimization, to appear in *Optimization Methods and Software.*
[2] B. Addis and S. Leyffer, A Trust-Region Algorithm for Global Optimization, submitted, available at `http://www.optimization-online.org/DB_HTML/2004/08/927.html`, Preprint ANL/MCS-P1190-0804, MCS Division, Argonne National Laboratory, Argonne, IL, 2004.
[3] The Cambridge Cluster Database, *http://www-wales.ch.cam.ac.uk/CCD.html*.
[4] D. M. Deaven, N. Tit, J. R. Morris, K. M. Ho, Structural optimization of Lennard-Jones clusters by a genetic algorithm, *Chemical Physics Letters* 256, 195–200 (1996).
[5] J. P. K. Doye, R. H.Leary, M. Locatelli and F. Schoen, The global optimization of Morse clusters by potential transformations, *INFORMS J. Computing*, 16, 371-379 (2004)
[6] Hartke B., Global cluster geometry optimization by a phenotype algorithm with niches: location of elusive minima, and low-order scaling with cluster size, *J.Comp.Chem.*, 20, 1752 (1999)
[7] C. Khompatraporn, Z.B. Zabinsky, Stopping and restarting strategy for stochastic adaptive search algorithms in global optimization, submitted
[8] R. H. Leary and J. P. K. Doye, Tethrahedral global minimum for the 98-atom Lennard-Jones cluster, *Phys. Rev. E*, R6320-R6322 (1999)
[9] R. H. Leary, Global optimization on funneling landscapes, *J. Global Optim.*, 18, 367–383, (2000).
[10] J. Lee, H. A. Scheraga, S. Rackovsky, New optimization method for conformational energy calculations on polypeptides: Conformational space annealing, *Journal of Computational Chemistry* 18(9), 1222-1232 (1997).
[11] J. Lee, I. H. Lee, J. Lee, Unbiased global optimization of Lennard-Jones clusters for $N \leq 201$ using the conformational space annealing method, *Physical Review Letters* 91(8): 080201/1-4 (2003).
[12] D.Liu, J.Nocedal, On the Limited Memory BFGS method for large scale optimization, *Mathematical Programming* B, 45, 503-528 (1989)
[13] M.Locatelli, F.Schoen, Efficient algorithms for large scale global optimization: Lennard-Jones clusters, *Computational Optimization and Applications*, 26, 173-190 (2003).
[14] M.Locatelli, On the multilevel structure of global optimization problems, to appear in *Computational Optimization and Applications*
[15] C.Lavor, N.Maculan, A function to test methods applied to global optimization of potential energy of molecules, *Numerical Algorithms*, 35, 287-300 (2004)
[16] C. Roberts, R. L. Johnston, N. T. Wilson, A genetic algorithm for the structural optimization of Morse clusters, *Theoretical Chemistry Accounts* 104, 123–130 (2000).
[17] Schwefel, H. P., *Numerical Optimization of Computer Models*, J. Wiley & Sons, Chicester (1981)
[18] Törn, A. and Žilinskas, A. *Global Optimization*, Lecture Notes in Computer Sciences, Springer-Verlag, Berlin, (1989)
[19] D. J. Wales, *Energy Landscapes with Applications to Clusters, Biomolecules and Glasses*, Cambridge University Press, Cambridge, (2003).
[20] D. J. Wales and J. P. K. Doye, Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms, *J. Phys. Chem. A*, 101, 5111–5116, (1997).
[21] D. J. Wales and H. A. Scheraga, Global Optimization of Clusters, Crystals and Biomolecules, *Science,* 285, 1368-1372 (1999).

Dip. Informatica - Università di Torino (Italy)
E-mail address: grosso@di.unito.it

Dip. Informatica - Università di Torino (Italy)
E-mail address: locatell@di.unito.it

Dip. Sistemi e Informatica - Università di Firenze (Italy)
E-mail address: schoen@ing.unifi.it