大 学 化 学
Univ. Chem. **2025,** *40* (9), 19

•专题•　　doi: 10.12461/PKU.DXHX202408095　　www.dxhx.pku.edu.cn

改进的模拟退火算法预测有机化合物分子式

陈晓东*，张玉敏
吉林大学化学学院，长春 130012

摘要：模拟退火算法是人工智能组合优化算法，在此算法的基础上，提出了一种改进的模拟退火算法，用于预测有机化合物分子式。算法开始，设计使用遗传算法计算种群各个体的适应度函数值，从中选择最优个体作为模拟退火算法初始解。然后在这个初始解的基础上，随机扰动生成新解，并计算其适应度函数值。若适应度函数值的增量小于等于零，则接受新解，否则按Metropolis准则判断是否接受新解。随着退火温度的缓慢降低，依据算法终止条件判断是否搜索到全局最优解。实验证明，该算法提高了搜索到全局最优解的成功率。将其用于预测有机化合物分子式时，其适应度函数收敛性明显优于经典模拟退火算法。

关键词：模拟退火算法；遗传算法；组合优化；质量分数；分子式
中图分类号：G64；O6；TP18

An Improved Simulated Annealing Algorithm for Predicting the Molecular Formulas of Organic Compounds

Xiaodong Chen *, Yumin Zhang
College of Chemistry, Jilin University, Changchun 130012, China.

Abstract: Simulated annealing algorithm is an artificial intelligence combinatorial optimization algorithm. Building upon the classic simulated annealing algorithm, we propose an enhanced version for predicting the molecular formulas of organic compounds. The algorithm begins by using a genetic algorithm to calculate the fitness values of individuals in the population, selecting the optimal individual as the initial solution for the simulated annealing process. Based on this initial solution, new solutions are generated through random perturbation, and their fitness values are calculated. If the change in fitness is less than or equal to zero, the new solution is accepted. Otherwise, the Metropolis criterion is applied to determine whether the new solution should be accepted. As the annealing temperature gradually decreases, the algorithm's termination condition is used to determine if the global optimal solution has been found. Experimental results show that this improved algorithm increases the success rate of finding the global optimal solution. When applied to predict the molecular formulas of organic compounds, it demonstrates significantly better convergence of the fitness function compared to the classical simulated annealing algorithm.

Key Words: Simulated annealing algorithm; Genetic algorithm; Combination optimization; Mass fraction; Molecular formulas

1 引言

在信息技术快速发展的今天，人工智能已悄然步入化学研究的各个领域。在化工、材料、生物、

收稿：2024-08-24；录用：2024-10-23；网络发表：2025-02-14
*通讯作者，Email: cxd@jlu.edu.cn
基金资助：2021年吉林大学实验技术项目(SYXM2021b001)；2021年吉林大学本科教学改革研究项目(2021XZC031)；2023年吉林大学本科教学改革研究重点项目(2023XZD039)

制药、环境等化学相关专业中引入人工智能相关课程、普及机器学习方法，是未来化学教学研究的一个方向[1]。人工智能领域重要的计算方法组合优化算法包含了模拟退火算法、遗传算法、蚁群算法、粒子群算法等多种优化算法，它们的优势在于，在给定约束条件下，能够寻找到多参数优化问题的全局最优解。经典模拟退火算法(Simulated Annealing，简称SA算法)的最早思想是由N. Metropolis等学者于20世纪50年代提出的[2]。1983年S. Kirkpatrick等成功将退火思想引入到组合优化领域[3]。该思想是源于蒙特卡洛迭代求解策略的一种随机寻找最优解的算法，是以一般组合优化问题与金属退火过程之间的相似性为出发点的[4]。SA算法从某一较高初始温度出发，利用Metropolis准则并模拟金属退火温度缓慢下降工艺而实现工件退火改性的目的，即通过该算法达到求解全局优化问题的目的[5]。SA算法是常用的优化算法之一，从理论上讲，该算法具有概率的全局优化性能。目前已被广泛地应用于神经网络、控制工程、信号处理、生产调度、机器学习等工程领域中[6–8]。

遗传算法(Genetic Algorithm) [9]是美国的J. H. Holland教授借鉴生物界自然选择和遗传进化机制在1975年首次提出的一种种群随机优化搜索算法[4]9–30 (中括号外数字代表引用书籍页码，下同)。本文在SA算法求初始解阶段，设计使用文献[10]遗传算法(以下简称遗传算法)，通过选择、交叉和变异等遗传操作，计算种群各个体的适应度函数值，从中选择最优个体作为SA算法初始解，避免随机产生初始解给SA算法搜索全局最优解带来的不确定性，提高了搜索到全局最优解的成功率。

一种新的有机化合物结构的确定一般需要利用质谱、元素分析、红外光谱及核磁共振氢谱和碳谱进行表征分析。可利用元素分析仪对有机化合物中的碳、氢、氮、氧、硫等元素的质量分数进行定量分析。如果该有机化合物中还含有卤素、磷等非金属元素时，还需借助其他化学分析法和分析仪器，如用等离子发射光谱仪等，来确定其质量分数，进而确定其分子式[11]。利用本文提供的改进的模拟退火算法，可在一定程度上简化上述过程，提高工作效率和节约成本。即在已知有机化合物分子量和元素种类的情况下，就可以预测该有机化合物的分子式。目前该方法可预测分子式中的元素种类最多达6种，每种元素数量范围从0–20个。因此，该算法在有机化合物结构分析方面有广阔的应用前景。

2 改进的模拟退火算法设计

在合金的热处理工艺中，为了提高金属材料的机械性能，将金属材料加热到一定温度，经保温，然后缓慢冷却的过程，被称之为退火。在上述退火过程中，金属原子随环境温度的变化重新排列而达到最低能量的稳定状态[12]，也就是达到了组合优化问题的最优解状态。

在退火过程中，系统的能量遵循Boltzmann概率分布规律，即系统依据概率$p(E)$处于能量为E的热平衡状态：

$$p(E) = \exp(-\delta E/kT) \tag{1}$$

(1)式中E为能量，T为温度，k为Boltzmann常数，δE为能量的变化量。该式说明随着环境温度T的降低，系统处于高能E状态的概率逐渐减小，这与金属退火规律相一致[13]。

SA算法首先要确定一个能量函数(适应度函数)，求解最优化问题一般通过Metropolis准则和模拟退火两个过程的协同来实现。Metropolis准则是在某一给定温度T的情况下，对解的状态空间进行随机抽样，当能量降低即$\delta E \le 0$时，接受当前状态；当能量升高即$\delta E > 0$时，则根据概率p有条件地接受当前解[13]。

$$p = \begin{cases} 1 & \text{if } \delta E \le 0 (f(x_2) \le f(x_1)) \\ \exp\left(-\dfrac{f(x_2) - f(x_1)}{kT}\right) & \text{if } \delta E > 0 (f(x_2) > f(x_1)) \end{cases} \tag{2}$$

(2)式中$f(x_1)$为当前解适应度函数，$f(x_2)$为新解适应度函数，k是Boltzmann常数，T是当前温度，能量变化量$\delta E = f(x_2) - f(x_1)$。

在新的温度条件下，不断地进行Metropolis准则抽样过程，直到Metropolis准则抽样和退火过程满足收敛条件为止。这种计算过程与金属退火工艺类似，即必须缓慢降温，才能使金属原子在每一温度下都达到热平衡，最终趋于全局能量最低的基态，而不是位于局部极小点[13]。

SA算法是一种随机搜索算法，表现在算法初始化阶段，随机选取初始解；为了防止陷入局部最优，在新解求解阶段，新解的随机扰动产生等。本文将一种改进的模拟退火算法(Improved Simulated Annealing Algorithm，简称ISAA算法)用于有机物分子式的预测，在算法初始化阶段，利用遗传算法群体搜索的优越性求取SA算法初始解，以改善SA算法搜索全局最优解时的盲目性，增加算法获得全局最优解的可能性和收敛速度。ISAA算法流程图如图1所示。

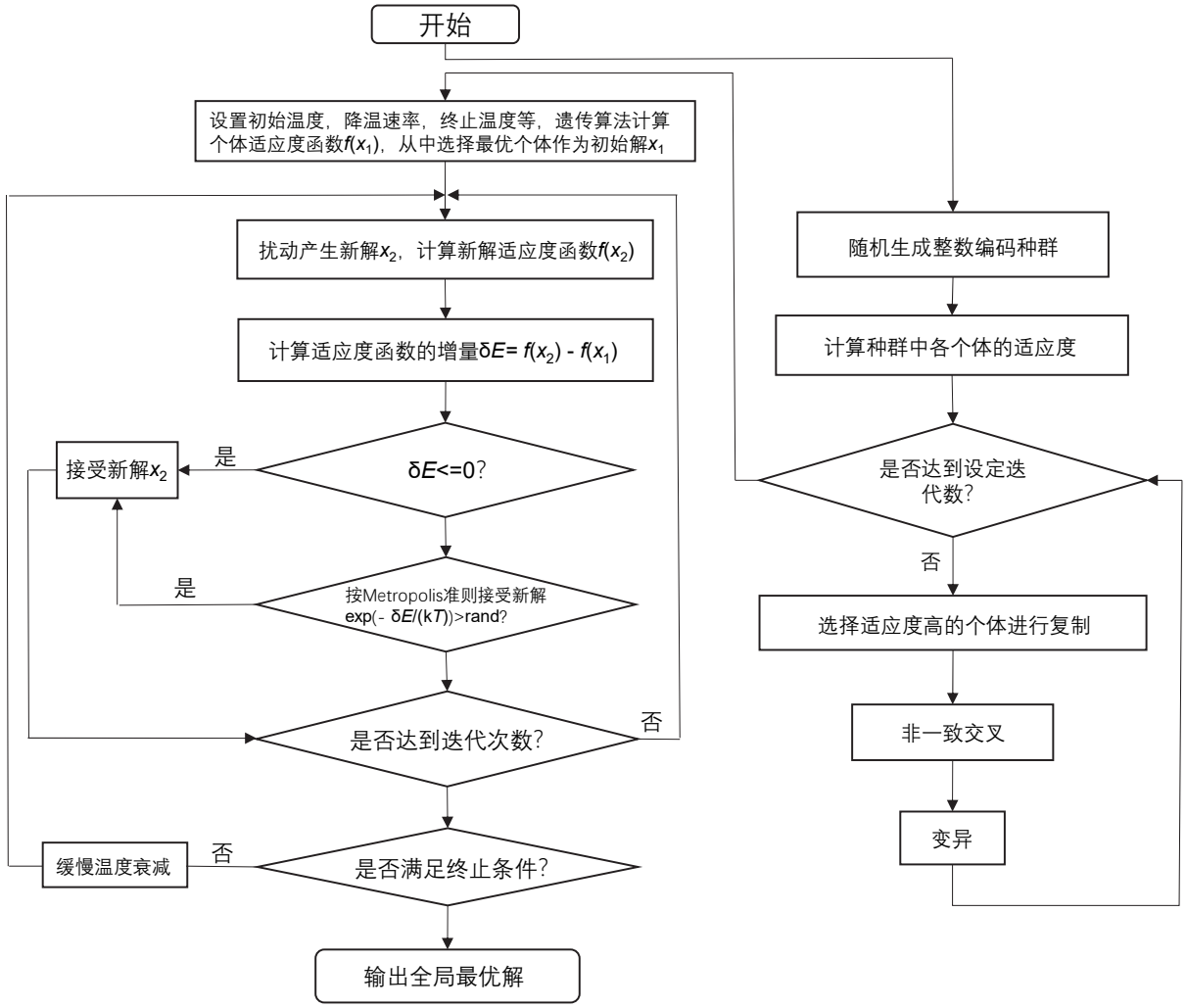

图1　ISAA算法流程图

该算法具体流程如下：

(1) 初始化：取初始温度T_0足够大，令$T = T_0$，使用遗传算法在整数空间计算种群中各个体的适应度函数值$f(x_1)$，从中选择最优个体作为SA算法初始解x_1。

(2) 对当前温度T，重复第(3)–(6)步骤。

(3) 对当前整数空间解随机扰动产生一个新整数空间解x_2，并计算其适应度函数值$f(x_2)$。

(4) 计算$f(x_2)$的增量$\mathrm{d}f = f(x_2) - f(x_1)$，其中$f(x_1)$为当前解适应度函数。

(5) 若$\mathrm{d}f < 0$，则接受x_2作为新的当前解，即$x_1 = x_2$；否则按Metropolis准则接受新解，计算$f(x_2)$的接受概率$\exp(-\mathrm{d}f/T)$，即随机产生(0, 1)区间上均匀分布的随机数rand，若$\exp(-\mathrm{d}f/T) > \mathrm{rand}$，接受$x_2$作为新当前解，即$x_1 = x_2$，否则保留当前解$x_1$。

(6) 如果满足终止条件，则输出当前解为全局最优解，算法结束；否则按衰减函数衰减T后返回第(2)步。

3 分子式的预测

整数规划问题是求决策变量取整数值的线性或非线性规划问题，是组合优化问题的一个分支[14]。预测有机化合物分子式可归结为整数规划问题，将接近目标分子量程度的函数作为适应度函数，描述为：

$$\begin{cases} \min \quad f(\mathrm{AtomC} * X_{\mathrm{C}} + \mathrm{AtomH} * X_{\mathrm{H}} + \mathrm{AtomN} * X_{\mathrm{N}} + \mathrm{AtomO} * X_{\mathrm{O}} + \mathrm{AtomS} * X_{\mathrm{S}} + \mathrm{AtomX} * X_{\mathrm{X}} - \mathrm{M}) \\ s.t. \quad X_{\min} \le X_i \le X_{\max} \quad (i = \mathrm{C, H, N, O, S, X}) \end{cases} \tag{3}$$

(3)式中f为适应度函数，AtomC、AtomH、AtomN、AtomO、AtomS、AtomX分别为碳、氢、氮、氧、硫、任意元素的精确原子量，M为目标分子量，决策变量X_{C}、X_{H}、X_{N}、X_{O}、X_{S}、Xx分别表示分子式中碳、氢、氮、氧、硫、任意元素的个数，决策变量$X_i \in [X_{\min}, X_{\max}] \in Z$ (i = C, H, N, O, S, X)，Z为非负整数空间，$X_{\min}$和$X_{\max}$为非负整数，每一个决策变量取一个值就构成解空间的一个解[14]。将预测有机化合物分子式问题归结为求(3)式适应度函数最小值。

针对有机化合物分子式预测这一整数规划问题，提出了ISAA算法。该算法将退火过程限制在整数空间，有效地提高了算法收敛速度。同时，使用遗传算法在SA算法求初始解阶段，通过选择、交叉和变异等遗传操作计算种群各个个体的适应度函数值，从中选择最优个体作为SA算法初始解，避免随机产生初始解给SA算法搜索全局最优解带来的不确定性，提高了搜索到全局最优解的成功率。

预测含有碳、氢、氮、氧、硫、铁六种不同元素的有机化合物分子式的算法基本步骤如下：

【步骤1】参数初始化。初始温度1e10，终止温度1e−30，各温度下的迭代次数800，降温速率0.99，决策变量$X_i \in [0, 20] \in Z$ (i = C, H, N, O, S, X)，Z为非负整数空间，并使用遗传算法计算种群中各个个体的适应度函数值$f(x_1)$，从中选择最优个体作为SA算法初始解x_1。

【步骤2】扰动产生新解x_2，并计算适应度函数值$f(x_2)$。

【步骤3】计算适应度函数的增量$\delta E = f(x_2) - f(x_1)$。

【步骤4】如果$\delta E \le 0$，则$x_1 = x_2$；否则按Metropolis准则接受新解，$\exp(-\delta E/(kT))$大于(0, 1)区间均匀分布随机数则接受新解，否则，抛弃新解，保留当前解。

【步骤5】是否达到设定的迭代次数，如果没有达到，则返回步骤2。

【步骤6】是否收敛于问题全局最优解，如满足，则输出全局最优解，否则返回步骤2。

【步骤7】输出全局最优解分子式。

依据本文提出的ISAA算法，采用Matlab 2020a语言编制程序。当测得的某一含有碳、氢、氮、氧、硫、铁六种元素有机化合物的分子量为264.0768时，设定上述每种元素的取值范围为[0, 20]，利用该程序计算得到该有机化合物分子式为$C_6H_{12}N_2O_4SFe$。ISAA算法和SA算法适应度函数收敛性比较详见图2，ISAA算法和SA算法收敛性比较详见表1。由图2可以看出，随着迭代次数的增加，ISSA算法适应度函数的收敛性优于SA算法。从表1可以看出，SA在预测分子式时，随着搜索范围的增加，收敛于目标分子量次数逐渐减小，而ISAA算法在各种搜索范围下均可收敛于目标分子量；随着搜索范围的增加，SA算法平均收敛代数逐渐增加，而ISAA算法在各种搜索范围下平均收敛代数增加幅度较SA算法少。

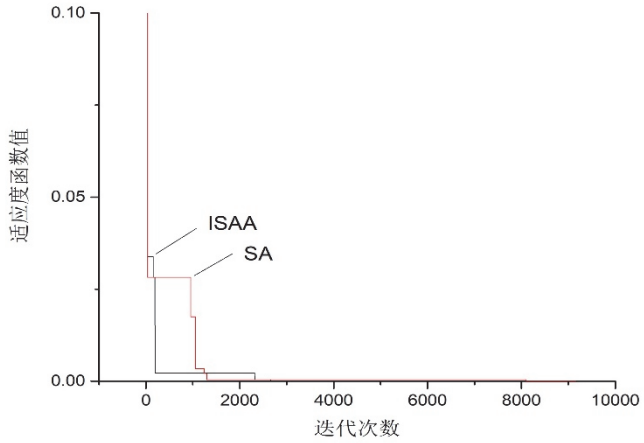

图2　ISAA算法和SA算法适应度函数收敛性比较

表1　ISAA算法和SA算法收敛性比较

搜索范围	SA算法收敛到全局最优解次数	ISAA算法收敛到全局最优解次数	SA算法平均收敛代数	ISAA算法平均收敛代数	目标分子式
$C_{10}H_{10}N_{10}O_{10}S_{10}Fe_{10}$	100	100	860	152	$C_6H_{12}N_2O_4SFe$
$C_{15}H_{15}N_{15}O_{15}S_{15}Fe_{15}$	98	100	2341	264	$C_6H_{12}N_2O_4SFe$
$C_{18}H_{18}N_{18}O_{18}S_{18}Fe_{18}$	95	100	3548	512	$C_6H_{12}N_2O_4SFe$
$C_{20}H_{20}N_{20}O_{20}S_{20}Fe_{20}$	90	100	4000	1847	$C_6H_{12}N_2O_4SFe$

4　结语

(1) 本文提出了预测有机化合物分子式的ISAA算法。使用遗传算法在SA算法求初始解阶段，通过选择、交叉和变异等遗传操作，计算种群各个体的适应度函数值，从中选择最优个体作为SA算法初始解，避免随机产生初始解给SA算法搜索全局最优解带来的不确定性，提高了搜索到全局最优解的成功率。

(2) 设计了接近目标分子量程度函数作为适应度函数。

(3) 将解空间约束在整数空间，避免算法搜索不必要的解空间，加快了算法收敛速度。

(4) 在已知有机化合物分子量及元素种类的情况下，利用本文的ISAA算法预测该有机化合物的分子式。

(5) 适当减少有机化合物中元素的种类，可以有效地提高该算法的收敛速度。

(6) 推广本文算法，可以预测元素种类在6个以下，各元素个数在20个以下的任意给定分子量物质的分子式。

(7) 本算法可以辅助学生在有机化合物元素分析、质谱分析以及二者协同分析过程中，快速确定有机化合物的分子式。

参 考 文 献

[1] 杜静, 于曦, 马骁飞, 赵温涛. 大学化学, **2024,** *39* (11), 65.

[2] 唐立山, 谢云, 尤矢勇, 罗祖华. 非数值并行算法(第一册). 模拟退火算法. 北京: 科学出版社, 1994: 22–55.

[3] Kirkpatrick, S.; Gelatt, C.; Vecchi, M. *Science* **1983,** *220* (4598), 671.

[4] 包子阳, 余继周, 杨杉. 智能优化算法及其MATLAB实例. 第1版. 北京: 电子工业出版社, 2016: 1–143.

[5] 李士勇, 李研. 智能优化算法原理与应用. 哈尔滨: 哈尔滨工业大学出版社, 2012: 40–46.

[6] Steinbrunn, M.; Moerkotte, G.; Kemper, A. *The VLDB Journal* **1997,** *6* (3), 191.

[7] 余娜, 何国荣, 李培东, 马驰. 计算机测量与控制, **2023,** *31* (11), 293.

[8] 黄智, 闵杰, 陈贵科, 饶志敏, 辛强, 赵寻. 天津大学学报(自然科学与工程技术版), **2024,** *57* (4), 374.

[9] 姚新, 陈国良. 计算机研究与发展, **1990,** No. 7, 1.

[10] Holland, J. H. *Adaptation in Natural Artificial Systems*; MIT Press: Cambridge, MA, USA, 1992; pp. 1–120.

[11] 陈晓东, 张玉敏. 计算机与应用化学, **2009,** *26* (1), 86.

[12] 陈晓东, 张玉敏, 徐跃. 计算机工程与应用, **2009,** *45* (27), 246.

[13] 张永贵, 陈明强, 李允, 胡受权. 西南石油学院学报, **1997,** *19* (3), 1.

[14] 陈晓东, 张玉敏. 计算机应用, **2009,** *29* (增刊), 165.