Bioorganic & Medicinal Chemistry Letters 22 (2012) 4540–4545

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

CCLab—a multi-objective genetic algorithm based combinatorial library design software and an application for histone deacetylase inhibitor design

Guanghua Fang [a,†], Mengzhu Xue [a,†], Mingbo Su [b], Dingyu Hu [a], Yanlian Li [a], Bing Xiong [a,*], Lanping Ma [a], Tao Meng [a], Yuelei Chen [a], Jingya Li [b], Jia Li [b,*], Jingkang Shen [a,*]

[a] State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Zhangjiang Hi-Tech Park, Pudong, Shanghai 201203, China
[b] National Drug Screening Center, 189 Guoshoujing Road, Zhangjiang Hi-Tech Park, Pudong, Shanghai 201203, China

ARTICLE INFO

Article history:
Received 20 March 2012
Revised 30 May 2012
Accepted 31 May 2012
Available online 7 June 2012

Keywords:
Combinatorial library design
Multi-objective optimization
Histone deactetylase
Inhibitor

ABSTRACT

The introduction of the multi-objective optimization has dramatically changed the virtual combinatorial library design, which can consider many objectives simultaneously, such as synthesis cost and drug-likeness, thus may increase positive rates of biological active compounds. Here we described a software called CCLab (Combinatorial Chemistry *Lab*oratory) for combinatorial library design based on the multi-objective genetic algorithm. Tests of the convergence ability and the ratio to re-take the building blocks in the reference library were conducted to assess the software in silico, and then it was applied to a real case of designing a 5 × 6 HDAC inhibitor library. Sixteen compounds in the resulted library were synthesized, and the histone deactetylase (HDAC) enzymatic assays proved that 14 compounds showed inhibitory ratios more than 50% against tested 3 HDAC enzymes at concentration of 20 μg/mL, with IC_{50} values of 3 compounds comparable to SAHA. These results demonstrated that the CCLab software could enhance the hit rates of the designed library and would be beneficial for medicinal chemists to design focused library in drug development (the software can be downloaded at: http://202.127.30.184:8080/drugdesign.html).

In the past decades, the advent of combinatorial chemistry has dramatically changed the drug discovery process, making it possible to perform a parallel synthesis a large number of chemical compounds for bioactivity assays.[1–4] However, it is not economic to synthesize a fully enumerated library, especially with an abundance of available building blocks.[5–8] In the meantime, although some leads have been identified by this approach, many of them failed in the following pharmacokinetics evaluations.[9] Hence, virtual combinatorial library design was introduced in order to reduce synthesis cost and increase the hit rates by applying certain filters or constraints during the library design process.[10–13]

To support the evaluation of multiple properties that medicinal chemists were interested in, library design has evolved to apply the multi-objective optimization technology.[14] Multi-objective optimization is a strategy that considers a number of objectives simultaneously during the library design phase, and ultimately yields a population of multi-dimensional solutions, each of which balances the defined objectives. There are some privately-owned software products to support this tactic. Among them, work from Gillet et al. is noteworthy. Previously, they developed a genetic algorithm based program SELECT with a fitness function by combining several weighted objectives.[15] Soon after, they reported an updated program MoSELECT, which was based on a multi-objective genetic algorithm and a fitness function based on the Pareto algorithm.[16–18] These software programs took a multi-component sequent linking method to build the molecules, in which building blocks would be connected step by step according to the defined reaction sequence.

Here we described a combinatorial library design software CCLab (Combinatorial Chemistry *Lab*oratory), which utilizes the multi-objective genetic algorithm based upon Pareto evaluation combined with two fragment connection modes, namely sequent linking and simultaneous linking. In this package, both the synthesis feasibility and evaluation of 'drug-like' properties could be considered, and the program was organized in a more flexible pattern, convenient to incorporate other programs to evaluate custom properties later.

The CCLab package consists of five modules that communicate with each other but perform different functions. As concretely illustrated in Figure 1, the Input module parses parameters from the input file and passes them to other modules; the CCLib module is responsible for assembling the fragments into molecules and building libraries on line; the CCScore module evaluates multiple

* Corresponding authors. Tel.: +86 2150806600x5412; fax: +86 2150807088.
E-mail addresses: bxiong@mail.shcnc.ac.cn (B. Xiong), jli@mail.shcnc.ac.cn (J. Li), jkshen@mail.shcnc.ac.cn (J. Shen).
† These authors contributed equally to this work.

0960-894X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.bmcl.2012.05.123

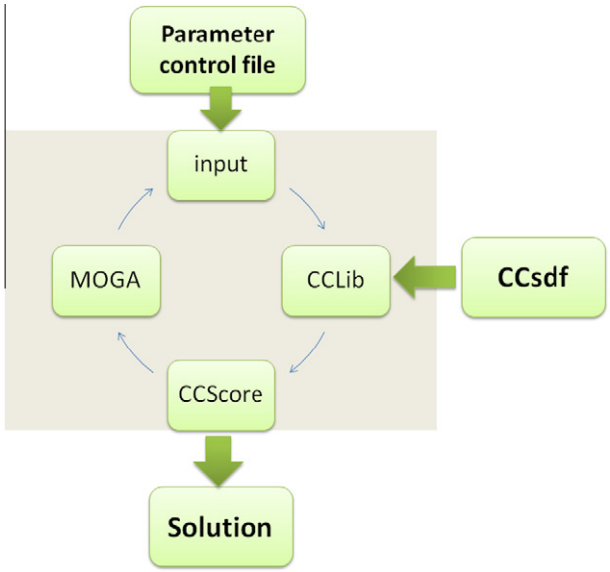


Figure 1. The flow chart of the functional modules in CCLab.

properties of the libraries, the MOGA module executes optimization with the multi-objective genetic algorithm and the main module integrates the modules above, performs the iterations of optimization and collects the output information during the library design process. The details of three important modules including CCLib, CCScore and MOGA were described as follows.

For the construction of a library, CCLib module provides two connection modes which are sufficient for most combinatorial library design. As shown in Figure 2, it can be summarized as a central scaffold with linking spots and functional groups to be linked, but differentiates in the mode of the functional groups connected to the scaffold simultaneously or successively. The program can automatically select proper building blocks from a large list to form a molecule by identifying the previously defined connection manner and reaction types.

In the properties calculation phase, CCScore module can conduct similarity calculation, diversity calculation, synthesis feasibility calculation and druglikeness calculation by calling external programs. Specifically, all the score expressions are formatted with the aim to minimize the parameters associated with the objectives. In details, the library similarity is to assess the resemblance between compounds in a generated library and the reference library, which is summarized the minimum score of each compound with respect to any compound in reference library, then divided this summary with compound number of the generated library to scale the similarity in the range 0–1. Currently, the CCScore can calculate two kinds of similarity, one is 2D fingerprint similarity and the other is 2D pharmacophore similarity calculated with ChemAxon program GenerateMD. The calculation of diversity of a library is implemented to average the dissimilarity of each pairwise compounds in the generated library. As for the drug-likeness calculation, a statistics-based method called *Z* score was adopted to calibrate it. First, some physico-chemical properties to describe drug-likeness were calculated for compounds in the reference and generated libraries. Secondly, normal distribution was modeled for each property *k* derived from the reference library and the mean μ_k with the standard deviation σ_k was obtained. It should be noted that the normal distribution may not fulfill all situations, and the users can modify this easily with little python programming skills. Thirdly, the $Z\ score_k^i$ value of each compound *i* in a generated library would be calculated by the following rules:

$$Z\ \text{score}_k^i = \begin{cases} 0 \\ \frac{|p_k^i - \mu_k|}{\sigma_k} \end{cases} \quad (1)$$

if the property p_k of this compound *i* is located within the range $\mu_k \pm 2\sigma_k$, the $Z\ score_k^i$ is 0; otherwise the $Z\ score_k^i$ would be calculated according to the formula above. Lastly, the sum of $Z\ score_k^i$ values of all the compounds in the generated library was taken as the property *k* derived penalty score of the generated library. The synthesis feasibility of the generated library was evaluated by an in-house program, which uses the machine learning method to correlate compound's synthesis feasibility with its properties, including mass, logP, aromatic atom number, aromatic ring number, asymmetric atom number, carbon ring number, fused aliphatic ring

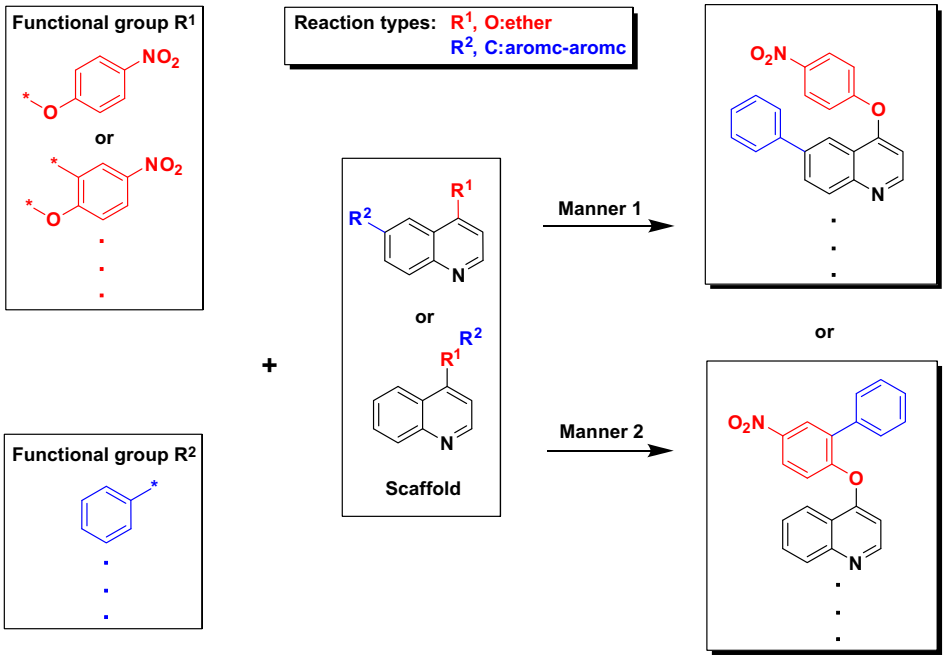


Figure 2. The 'scaffold-functional groups' connection modes used in the program.

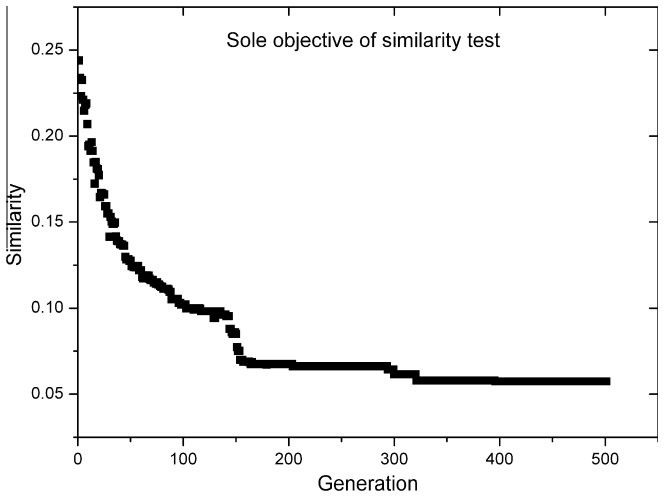


Figure 3. The result of fingerprint similarity sole-objective tests. The tests were carried out with the optimal parameters and the similarity value in the best Pareto ranking of every generation in each test were averaged.

number, fused aromatic ring number, hetero-ring number, tautomer number and wiener index (details in Supplementary data). Then the average synthesis feasibility score of each compound was defined as the score of the generated library.

MOGA module is comprised by elementary components of a standard genetic algorithm, mainly an initialization operator, a fitness operator, a selection operator, a crossover operator and a mutation operator.[19,20] And the multi-objective evaluation method Pareto ranking was adopted (details are provided in Supplementary data). [21] The fitness values of individuals are determined using the concept of dominance where an individual is non-dominated if a score of it in at least one of its objectives is not worse than others', finally leading to a Pareto surface filled with all non-dominated solutions that are considered equivalent. And the niche sharing method was applied into the fitness evaluation to distribute solutions into different clusters or niches, and the fitness of solutions in the same niche would be penalized. In general, MOGA would generate a group of non-dominated solutions with compromised consideration of multi-objectives and clustered with a given niche radius to decrease the probability of similar individuals.

In summary, by comparing with MoSelect, CCLab inherit the principles of the multi-objective genetic algorithm and Pareto ranking method from MoSELECT. While the linking modes, evaluations of various properties and the synthesis feasibility were largely different from MoSELECT. Besides, our software can be extended easily to implement other methods to calculate druglike properties.

To assess the software utility, we first conducted in silico evaluations. All the functional groups were obtained from our previous work,[22] and the scaffold for evaluation was chosen with no target

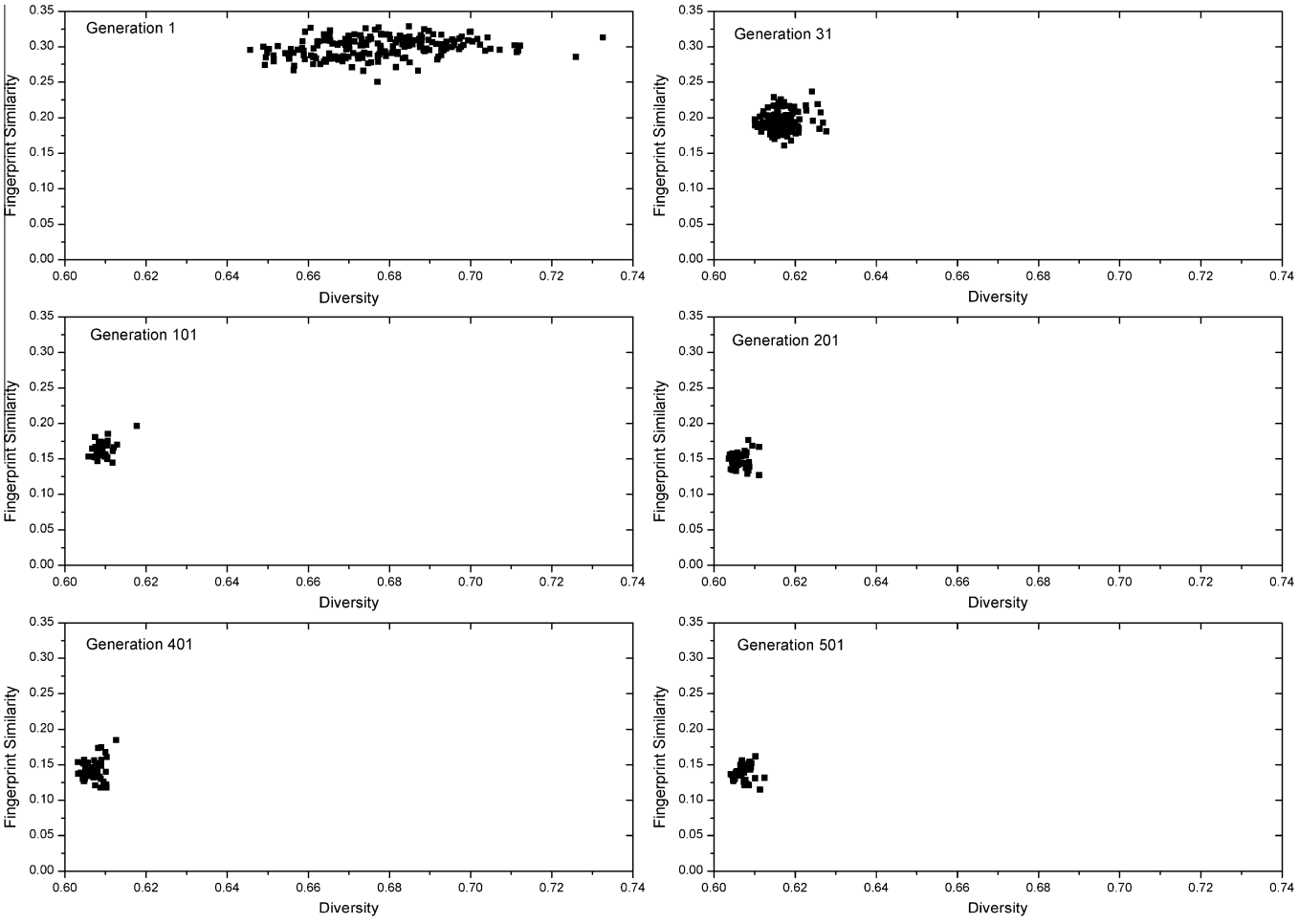


Figure 4. The result of a two-objective tests consisting of the fingerprint similarity and the fingerprint diversity. The similarity and diversity values were collected from the data in the best Pareto ranking of generation 1, 31, 101, 201, 401 and 501. Noting that the minimization is defined as the optimization trend of all the objective functions, the lower left space represents a superior region.

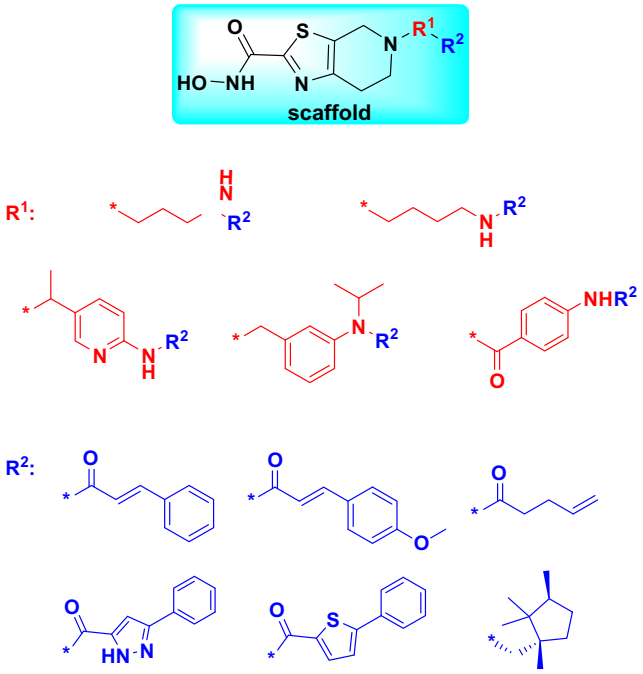


Figure 5. The composition of the final selected library. It is defined as the form of scaffold–R^1–R^2, and the fragments are listed respectively with stars showing as the point to be linked.

bias, on which reaction types of the two attachment sites were defined manually. The reference library has the size of 10 × 10, and each 10 building blocks of two reaction sites were randomly selected from two 1000 building block pools respectively. During the in silico test, we first optimized the MOGA parameters in the software by using the fingerprint similarity as the solo objective to be inspected. Then various combinations of parameters were carried out for the optimization and only one parameter was allowed to change in the process of one test run.

For each parameter, an optimal value would be picked out from a group of parallel tests with other parameters fixed. The assessment was executed by 2D fingerprint similarity sole-objective tests. Figure 3 shows the average result of the similarity sole-objective test. Generally speaking, the convergence speed of the program is superior to MoSELECT, and usually gets to an obvious convergence at about the 300th generation. The ultimate fingerprint similarity could converge to about 0.05, leading to resulted libraries of 95% similarity. In addition, to libraries of a scaffold with two attachment sites, 8 fragments of R^1 and 7 fragments of R^2 in the reference could be found simultaneously in resulted libraries of the last generation.

Consequently in the two-objective tests, each property would be chosen and then combined with the fingerprint similarity. The result of a group consisting of the fingerprint similarity and the fingerprint diversity is illustrated in Figure 4 as an example. Compared with the sole-objective test, it still shows a similar convergence, resulting in that the fingerprint similarity initialized at lib_sim 0.3 but stabilizes at 0.125 with a slight fluctuation step by step. Similarly, 5 fragments of R^1 and 6 fragments of R^2 in the reference are found simultaneously in resulted library of the last generation (The optimized parameters were listed in Supplementary data Table S3).

In addition to the in silico tests, a real application was conducted to verify the utility of the CCLab by designing a focused library of HDAC inhibitors. In present work, 49 reported HDAC inhibitors were collected as the reference library. Among them, 27 HDAC inhibitors mostly in clinical phase were chosen as representatives for calculating the drug-like properties to enable the designed library having good PD/PK properties (The ligand structures were listed in Table S1 and S2 in Supplementary data.).

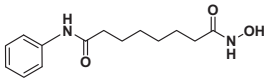
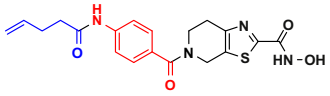
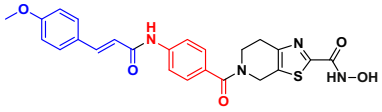
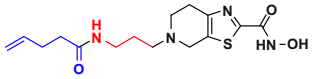
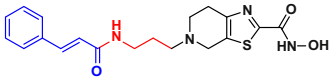
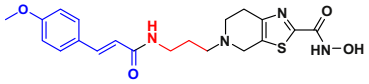
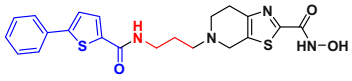
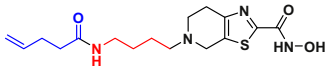
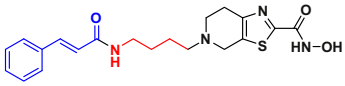
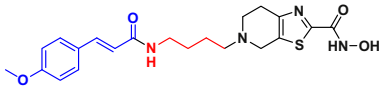
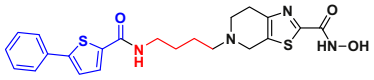
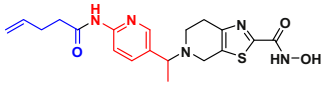
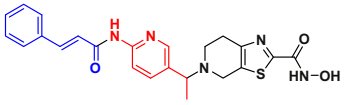
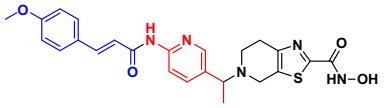
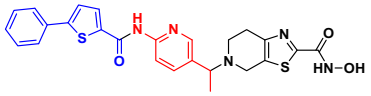
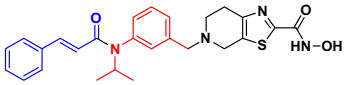
In this application, 7 objectives were considered in the design of HDAC inhibitor library, including fingerprint similarity, pharmacophore diversity, synthetic feasibility, log*P*, mass, rotatable bond number and polar surface area. Other MOGA parameters were set to the optimal values as identified by parameter optimization phase, and the size of resulted libraries was set to 5 × 6. The scaffold was designed based on a class of HDAC inhibitors reported in a patent.[23] The final objective values were listed in Table S4 in Supplementary data. Since the MOGA is stochastic modeling algorithm, to better explore the solution space, CCLab was run 10 times to generate more virtual libraries, then a Pareto ranking based python script was used to rank them. Finally, the library belonged to the best pareto ranking was illustrated in Figure 5 and advised to synthesis. Similar to many known HDAC inhibitors, the scaffold defined by us also has a hydroxamic acid group, which is intended to form a chelate interaction with the Zn^{2+} atom in the binding site of HDAC proteins.[24–27] But the designed R^1 and R^2 are very different from reported HDAC, which may introduce novelty for HDAC inhibitor development.

Finally, 16 compounds in the library were synthesized[28] limited by the commercially available reagents and then were assessed by inhibitory activity assays against HDAC1, 3, and 6.[29] As listed in the Table 1, the compounds can be divided into 4 groups by different R^1 fragments. Compounds **7a**–**7c** were attributed to Group A and **19a**–**19d** belonged to Group D, in which R1 fragments were both aromatic, whereas **11a**–**11d** belonged to Group B and **15a**–**15d** were divided to Group C, in which R^1 fragments were both aliphatic. Initial activity assays were indicated that, with drug SAHA as the control compound, 14 compounds showed inhibitory ratios more than 50% at the concentration of 20 μg/mL against 3 HDAC enzymes. From IC_{50} values, 3 compounds were comparable to SAHA against HDAC6. And in details, compounds showed better activities against HDAC enzymes when they containing aromatic rings in both R^1 and R^2 building blocks. It is also noted that the compounds **11**, **15**, 19 and **23** all contain an amine group. By comparing with compounds **7**, it is found that the linking part with amide group may be better for HDAC6 interactions. And further follow-up development of this series of HDAC inhibitors will be reported elsewhere.

Multi-objective optimization has greatly changed the virtual combinatorial library design, which can consider many properties simultaneously. In this report, we developed a multi-objective genetic algorithm based combinatorial library design software package CCLab (*C*ombinatorial *C*hemistry *Lab*oratory). The software incorporates molecular similarity, synthesis feasibility and 'lead-like' properties into the multi-objective evaluation, and uses the genetic algorithm to implement the optimization.

In silico tests using an in house training set were carried out to assess the software. The results indicated the software can converge in the reasonable time scale about 5–10 h on an Intel XEON 2.8G processor for 10 times run of HDAC inhibitor design. From these tests, it was found that the CCLab can find most of the pre-inserted fragments. Furthermore, the software was applied for design of a HDAC inhibitor combinatorial library. Finally, 16 compounds were synthesized and evaluated by bioactivity assays. Among them, 14 compounds showed moderate inhibitory potencies against tested 3 HDAC enzymes, some were exhibited selectivity against HDAC6, and 3 compounds have the IC_{50} values comparable to the positive control marketed drug SAHA. Clearly, the CCLab software can enhance the hit rates and would be beneficial for combinatorial library design.

Table 1
The activities of the 16 compounds against HDAC1, 3, and 6[a]

Codes	Structure	HDAC1 IC_{50} (μm)	HDAC3 IC_{50} (μm)	HDAC6 IC_{50} (μm)
SAHA		0.18 ± 0.03	0.14 ± 0.02	0.12 ± 0.01
7a		9.70 ± 1.83	9.38 ± 1.03	0.39 ± 0.07
7b		22.5 ± 6.2	16.07 ± 3.25	0.45 ± 0.04
7c		2.59 ± 0.27	2.66 ± 0.53	0.12 ± 0.03
11a		NA	NA	58.67 ± 13.39
11b		39.43 ± 14.25	13.48 ± 4.14	13.46 ± 1.73
11c		10.32 ± 2.18	6.62 ± 1.37	8.77 ± 1.55
11d		5.67 ± 0.97	5.09 ± 0.53	3.39 ± 0.68
15a		NA	NA	24.75 ± 9.06
15b		40.12 ± 8.83	26.39 ± 6.28	23.18 ± 3.19
15c		35.07 ± 6.96	34.12 ± 5.12	9.55 ± 0.91
15d		6.08 ± 1.88	4.89 ± 1.33	2.69 ± 0.40
19a		65.71 ± 8.49	19.20 ± 3.54	13.50 ± 2.38
19b		6.8 ± 1.76	4.80 ± 0.91	1.07 ± 0.05
19c		6.8 ± 0.54	5.7 ± 0.65	1.55 ± 0.31
19d		2.96 ± 0.99	1.94 ± 0.64	1.33 ± 0.32
23a		11.49 ± 1.60	2.19 ± 0.26	1.22 ± 0.33

[a] IC_{50} values were obtained if the inhibition rate was larger than 50%. NA, not determined. Red colored parts in compounds are R^1 group, while R^2 were colored as blue.

Acknowledgment

This work is supported by the State Key Laboratory of Drug Research. The authors gratefully acknowledge financial support from National Science & Technology Major Project "Key New Drug Creation and Manufacturing Program" of China (Grant No. 2009ZX09501-010), Program of Excellent Young Scientist of Chinese Academy of Sciences (Grant No. KSCX2-EW-Q-3-01) and

the National Natural Science Foundation of China (Grant 81072580 and 81102306).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bmcl.2012.05.123.

References and notes

1. Khosla, C. *Curr. Opin. Biotechnol.* **1996**, *7*, 219.
2. Fauchere, J.-L.; Boutin, J. A.; Henlin, J.-M.; Kucharczyk, N.; Ortuno, J.-C. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 43.
3. Hijfte, L. V.; Marciniak, G.; Froloff, N. *J. Chromatogr., B: Biomed. Sci. Appl.* **1999**, *725*, 3.
4. Williard, X.; Pop, I.; Horvath, D.; Bourel, L.; Melnyk, P.; Deprez, B.; Tartar, A. *Eur. J. Med. Chem.* **1996**, *31*, 87.
5. Furka, A. *Drug Discovery Today* **2002**, *7*, 1.
6. Gray, N. S. *Curr. Opin. Neurobiol.* **2001**, *11*, 608.
7. Lobanov, V. S. *Trends Biotechnol.* **2002**, *20*, 86.
8. Weber, L. *Drug Discovery Today* **2004**, *1*, 261.
9. Edwards, P. J. *Drug Discovery Today* **2009**, *14*, 108.
10. Leach, A. R.; Hann, M. M. *Drug Discovery Today* **2000**, *5*, 326.
11. Weber, L. *Curr. Opin. Chem. Biol.* **2000**, *4*, 295.
12. Li, J.; Murray, C. W.; Waszkowycz, B.; Young, S. C. *Drug Discovery Today* **1998**, *3*, 105.
13. Waldman, M.; Li, H.; Hassan, M. *J. Mol. Graphics Modell.* **2000**, *18*, 412.
14. Schneider, G. *Curr. Med. Chem.* **2002**, *9*, 2095.
15. Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, DVS *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169.
16. Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, DVS *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 375.
17. Gillet, V. J.; Willett, P.; Fleming, P. J.; Green, DVS *J. Mol. Graphics Modell.* **2002**, *20*, 491.
18. Wright, T.; Gillet, V. J.; Green, D. V. S.; Pickett, S. D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 381.
19. Sheridan, R. P.; SanFeliciano, S. G.; Kearsley, S. K. *J. Mol. Graphics Modell.* **2000**, *18*, 320.
20. Chen, G.; Zheng, S.; Luo, X.; Shen, J.; Zhu, W.; Liu, H.; Gui, C.; Zhang, J.; Zheng, M.; Puah, C.; Chen, K.; Jiang, H. *J. Comb. Chem.* **2005**, *7*, 398.
21. Konak, A.; Coit, D. W.; Smith, A. E. *Reliab. Eng. Syst. Saf.* **2006**, *91*, 992.
22. Yan, B. B.; Xue, M. Z.; Xiong, B.; Liu, K.; Hu, D. Y.; Shen, J. K. *Acta. Pharmacol. Sinica* **2009**, *30*, 251.
23. Miller, T. A.; Witter, D. J.; Belvedere, S. WO2005034880, 2005.
24. Kazantsev, A. G.; Thompson, L. M. *Nat. Rev. Drug Disc.* **2008**, *7*, 854.
25. Minucci, S.; Pelicci, P. G. *Nature* **2006**, *6*, 38.
26. Monneret, C. *Eur. J. Med. Chem.* **2005**, *40*, 1.
27. Paris, M.; Porcelloni, M.; Binaschi, M.; Fattori, D. *J. Med. Chem.* **2008**, *51*, 1505.
28. The reagents (chemicals) were purchased from Lancaster, Acros, and Shanghai Chemical Reagent Co. and used without further purification. Analytical thin-layer chromatography (TLC) was HSGF 254 (150–200 μm thickness; Yantai Huiyou Co., China). Nuclear magnetic resonance (NMR) spectroscopy was performed on a Bruker AMX-400 and AMX-300 NMR (IS as TMS). Chemical shifts were reported in parts per million (ppm, δ) downfield from tetramethylsilane. Proton coupling patterns were described as singlet (s), doublet (d), triplet (t), quartet (q), multiplet (m), and broad (br). Low- and high-resolution mass spectra (LRMS and HRMS) were given with electric, electrospray, and matrix-assisted laser desorption ionization (EI, ESI, and MALDI) produced by a Finnigan MAT-95, LCQ-DECA spectrometer and IonSpec 4.7 T. The purity of final compounds was assessed by the analytical HPLC method and found to be >95%. An Agilent 1100 series HPLC with an Agilent Zorbax Exlipse SB-C18 (25–4.6 mm, 5 μm particle sizes) reversed-phase column was used for analytical HPLC analyses. The elution buffer was an A/B gradient, where A = H_2O and B = CH_3OH. All reactions were carried out under dry and inert condition unless otherwise stated. Experimental details for target compounds, some intermediates and other compounds are included in the Supplementary data.
29. The fluorogenic histone deacetylase assay was carried out as Wegener, D. et al. reported[30]. Human His6-tagged and GST-fusion HDAC proteins were expressed in insect High5 cells using a baculoviral expression system, and purified using Ni-NTA (QIAGEN). The deacetylase activity of rhHDACs (recombinant human HDACs) 1 and 3 was assayed with a HDAC substrate (Ac-Lys-Tyr-Lys(ε-acetyl)-AMC), and HDAC6 was assayed with another HDAC substrate(Boc-Lys(ε-acetyl)-AMC). The total HDAC assay volume was 25 μl and all the assay components were diluted in Hepes buffer (25 mM Hepes, 137 mM NaCl, 2.7 mM KCl and 4.9 mM MgCl2, pH 8.0). The reaction was carried out in black 384-well plates (OptiPlateTM-384F, PerkinElmer). In brief, the HDAC assay mixture contained HDAC substrate (5–50 μM, 5 μl), rhHDAC isoforms (20–200 nM) and inhibitor (1 μl). SAHA was used as the positive control for all the HDACs assay. The negative controls contained neither enzyme nor inhibitor. The incubation time for the HDAC6 assay is 3 h, and for HDAC1, 3 is 24 h at room temperature. The reaction was quenched with the 25 μl trypsin addition (diluted to final concentration 0.3125%). Then the plates were incubated for 30 min at room temperature to allow the fluorescence signal to develop. At last, the fluorescence generated was monitored at wavelengths 355 nm (excitation) and 460 nm (emission) using Envision (PerkinElmer). Compounds were tested in 8-dose in duplicate with two or three fold serial dilution. And the IC_{50} data was calculated using the software GraphPad Prism, and chosen the equation 'sigmoidal dose–response (variable slope)' for curve fitting.
30. Wegener, D.; Wirsching, F.; Riester, D.; Schwienhorst, A. *Chem. Biol.* **2003**, *10*, 61.