RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2019, **9**, 3912

LSA: a local-weighted structural alignment tool for pharmaceutical virtual screening†

Xiuming Li,‡[a] Xin Yan,‡*[a] Yuedong Yang,[c] Qiong Gu,[a] Huihao Zhou,[a] Yunfei Du,[c] Yutong Lu,[c] Jielou Liao[d] and Jun Xu*[ab]

Similar structures having similar activities is a dogma for identifying new functional molecules. However, it is not rare that a minor structural change can cause a significant activity change. Methods to measure the molecular similarity can be classified into two categories of overall three-dimensional shape based methods and local substructure based methods. The former states the relation between overall similarity and activity, and is represented by conventional similarity algorithms. The latter states the relation between local substructure and activity, and is represented by conventional substructure match algorithms. Practically, the similarity of two molecules with similar activity depends on the contributions from both overall similarity and local substructure match. We report a new tool termed as a local-weighted structural alignment (LSA) tool for pharmaceutical virtual screening, which computes the similarity of two molecular structures by considering the contributions of both overall similarity and local substructure match. LSA consists of three steps: (1) mapping a common substructure between two molecular topological structures; (2) superimposing two three-dimensional molecular structures with substructure focus; (3) computing the similarity score based on superimposing. LSA has been validated with 102 testing compound libraries from DUD-E collection with the average AUC (the area under a receiver-operating characteristic curve) value of 0.82 and an average $EF^{1\%}$ (the enrichment factor at top 1%) of 27.0, which had consistently better performance than conventional approaches. LSA is implemented in C++ and run on Linux and Windows systems.

Received 28th October 2018
Accepted 23rd January 2019

DOI: 10.1039/c8ra08915a

rsc.li/rsc-advances

Introduction

Ligand-based virtual drug lead screening[1] is based on the principle of "similar structures having similar activities".[2] There are many methods to measure the similarity of two molecular structures. These methods can be classified into two categories: (1) overall three-dimensional shape based methods, such as ROCS[3] and WEGA,[4] and (2) local substructure based methods,[5] such as atom-pairs,[6] ECFP,[7] or substructure search methods.[8] The former uses the relation between overall steric similarity and activity regardless of covalent connectivity. The latter uses the relation between substructure (local covalent connectivity) and activity regardless of global shape.[9] Practically, the similarity of two molecules with similar activity depends on both overall and local similarity factors,[10] but also global shape and local substructures. There is not yet a similarity method that can combine both overall and local similarity factors. Therefore, the similarity measured by shape based methods cannot result in consistent similarity activity relations;[11] and the substructure or atom-pair search algorithms cannot satisfy scientists in discovering novel lead compounds or elucidating activity–substructure relations.[12]

In medicinal chemistry, functional groups (substructures) at a molecule do not contribute to the activity equally. One substructure[13] can be significantly more important than the other substructures, and is termed as a privileged substructure (or fragment).[14] Fig. 1 shows an HDAC (histone deacetylase)[15] inhibitor and its privileged substructure (highlighted in red circle). This substructure is the core substructure because a pan HDAC inhibitor must have a chelator "warhead" binding Zn^{2+} ion. Without this core substructure, the agent will not be active regardless of how the rest of the molecule is similar to an HDAC inhibitor. A substructure match algorithm (such as

[a]Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou 510006, China. E-mail: junxu@biochemomes.com
[b]School of Computer Science & Technology, Wuyi University, 99 Yingbin Road, Jiangmen 529020, China
[c]National Supercomputer Center in Guangzhou, School of Data and Computer Science, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou 510006, China
[d]Department of Chemical Physics, University of Science and Technology of China, Jinzhai Road 96, Hefei 230026, China

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8ra08915a

‡ Equal contributors.

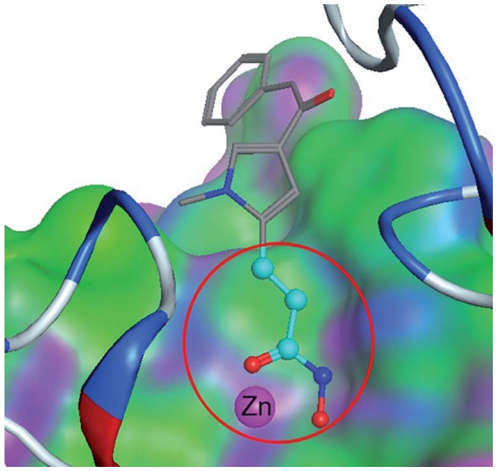

Fig. 1 A core substructure of an HDAC inhibitor. The core substructure (highlighted in red circle) is for an HDAC inhibitor (CHEMBL275089). The core substructure is the chelator "warhead" binding Zn^{2+} ion in the HDAC binding site. The rest of the molecular structure is for selective molecular recognition.

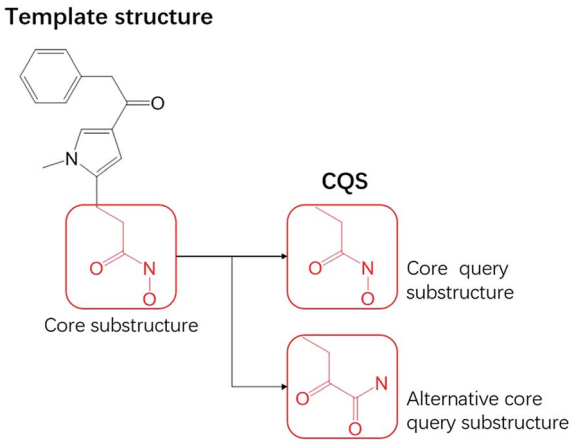

Fig. 2 The specification of substructures. The core substructure is specified from a template molecule which is as the core query substructure. One (or more) alternative core query substructure(s) is specified.

GMA[8]) can be employed to determine if a molecule is qualified for a potential HDAC inhibitor by checking the chelator[16] "warhead" existence in the molecule.

However, the rest of an HDAC inhibitor is still important and responsible for selectively binding to HDAC target (molecular recognition). The molecular recognition part of the HDAC inhibitor is associated with the overall molecular structure similarity,[17] which can be calculated through global shape comparison (three-dimensional structure superimposing). A molecular shape comparison algorithm can be used to predict the potency of a molecular being an HDAC inhibitor by calculating the overall similarity to a known HDAC inhibitor.

Therefore, LSA is reported to compute the similarity of two molecular structures by considering the contributions of both overall similarity and local substructure match.

LSA consists of the following main steps:

(1) Mapping a common substructure between two molecular topological structures.

(2) Superimposing two three-dimensional molecular structures with substructure focused. LSA will assign weights to atoms in the substructure mappings acquired from step (1) when superimposing.

(3) Computing the similarity score based on the superimposing using Tanimoto protocol.

Methods

Specifying core query substructure

A core query substructure (CQS) is a common substructure between two molecular topological structures, and can be derived from a template molecule (a compound with known activity).[18] A CQS represents topological features a hit candidate must possess. From a template structure, a user can specify more than one substructure (Fig. 2) for a CQS.

Mapping common substructure

By reference to GMA,[8] CQS are mapped from a template molecule (A) to a potential hit structure (B). If mapping $\boldsymbol{M}$ (CQS, A $\rightarrow$ B) $\neq \varnothing$, then $\boldsymbol{M}$ can have multiple mappings. Each mapped atom is marked as the more important atom than non-mapped atoms in molecules A and B.

Superimposing two steric structures with the substructure mappings

Restricted WEGA (rWEGA), a modified WEGA, was developed to conduct the conditional structural superimposing with the restrictions of the substructure mappings.

With such restriction, rWEGA will no longer treat every overlaid atom-pair equally while calculating steric structure similarity. The atoms in the atom-pairs of the mappings will be assigned with a weight w_a (if the atom in molecule A) or w_b (if the atom in molecule B) to address that these atoms are more important than other atoms regarding the contributions to the activity. The weights are computed with eqn (1) and (2).

$$w_a = \sqrt{\frac{n_A}{n_{AC}}} \quad (1)$$

$$w_b = \sqrt{\frac{n_B}{n_{BC}}} \quad (2)$$

where, n_A is the number of atoms in molecule A, n_{AC} is the number of atoms in the core substructure in molecule A, n_B is the number of atoms in molecule B, n_{BC} is the number of atoms in the core substructure in molecule B.

The LSA similarity scoring calculation in rWEGA is described in Algorithm 1.

Algorithm 1: LSA Similarity Scoring

```
input: A, B, CQS
do substructure mapping
    get M
i = 1, j = 1
if (M ≠ ∅) then
    while (i ≦ nc) do
        while (j ≦ nm) do
            calculate w_a, w_b
            move alignment center to the substructures center
            calculate V_AC, V_BC, V_AE, V_BE,
            calculate V_C, V_E
            calculate S_c(M_j), S_ec(M_j)
            calculate S (M_j)
        calculate S_i(M)
    calculate S(A, B)
    return S(A, B)
```

where, n_c is the number of conformations of B, n_m is the number of mappings in B, V_{AC} is the self-overlap volume of the core substructure in molecule A, V_{BC} is the self-overlap volume of the core substructure in molecule B, V_{AE} is the self-overlap volume of molecule A excluding its core substructure, V_{BE} is the self-overlap volume of molecule B excluding its core substructure, V_C is the overlap volume of the core substructure in molecule A and the core substructure in molecule B, and V_E is the overlap volume of molecule A and molecule B excluding the core substructures.

Let v_{ij} be the intersection weighted-Gaussian volume[4] of atom i and atom j. V_{AC}, V_{BC}, V_{AE}, V_{BE}, V_C, V_E are computed in eqn (3)–(8):

$$V_{AC} = \sum_{i\in AC, j\in AC} w_a w_a v_{ij} \tag{3}$$

$$V_{BC} = \sum_{i\in BC, j\in BC} w_b w_b v_{ij} \tag{4}$$

$$V_{AE} = \sum_{i\in AE, j\in AE} v_{ij} \tag{5}$$

$$V_{BE} = \sum_{i\in BE, j\in BE} v_{ij} \tag{6}$$

$$V_C = \sum_{i\in AC, j\in BC} w_a w_b v_{ij} \tag{7}$$

$$V_E = \sum_{i\in AE, j\in BE} v_{ij} \tag{8}$$

$S_c(M_j)$ is the core substructure similarity of A and B based on the jth mapping. $S_{ec}(M_j)$ is the structural similarity of A and B based on the jth mapping excluding the core substructures. $S(M_j)$ is the similarity of A and B based on the jth mapping. If $\boldsymbol{M}$ have multiple substructure mappings, take the maximum $S(\boldsymbol{M})$ as the similarity of A and B. $S_c(M_j)$, $S_{ec}(M_j)$, $S(M_j)$ and $S(\boldsymbol{M})$ are computed in eqn (9)–(12):

$$S_c(M_j) = \frac{V_C}{V_{AC} + V_{BC} - V_C} \tag{9}$$

$$S_{ec}(M_j) = \frac{V_E}{V_{AE} + V_{BE} - V_E} \tag{10}$$

$$S(M_j) = \sqrt{S_c(M_j) S_{ec}(M_j)} \tag{11}$$

$$S(\boldsymbol{M}) = \mathrm{Max}(S(M_j)), j \in 1\ldots n_m \tag{12}$$

If molecule B have multiple conformations, take the maximum as the final similarity of A and B. Let $S_i(\boldsymbol{M})$ be the similarity of A and B for the ith conformation of B, S(A, B) is the final similarity score of molecule A and B calculated from the values of $S_i(\boldsymbol{M})$ as shown in eqn (13).

$$S(\mathrm{A}, \mathrm{B}) = \mathrm{Max}(S_i(\boldsymbol{M})), i \in 1\ldots n_c \tag{13}$$

Method for validating LSA

The validation data were taken from the Directory of Useful Decoys collection[19] (DUD-E) which consists of 102 compound libraries, which are associated with 102 protein targets. Each targeted library has one template active compound, active and "decoy" compounds, and their chemical structures.§

In order to validate LSA, three-dimensional conformations of the compound structures in the libraries were generated by CAESAR[20] module in Discovery Studio (version 3.5) with the energy interval of 20 kcal mol^{-1}. The CQS were specified by reference to the common structure of "active" molecules derived from DUD-E.¶

AUC (the area under a receiver-operating characteristic curve) values and enrichment factors (EF) at the top x% ($x = 1$, 5, 10) are used to measure the performance of LSA when it is used in virtual pharmaceutical screening experiments. $EF^{x\%}$ is calculated:

$$EF^{x\%} = (TP^{x\%}/N^{\%}_{selected})/(N_{actives}/N_{total}) \quad (14)$$

where $TP^{x\%}$ and $N^{\%}_{selected}$ are the number of true positives and the number of selected candidates at the top x% of the screening library. $N_{actives}$ and N_{total} are the number of active compounds and the total number of the screening library. $EF^{x\%}$ is the fraction of active molecules at the cutoff x% of the database screened, which can represent how efficiently known active molecules can be differentiated compared to the random selections.

Results

The results of virtual screening 102 targeted libraries

102 targeted libraries were virtually screened using template structures with specified core substructures. It costs about 20 minutes to screen every 10 000 molecules (each with 50 conformations). The virtual screening performances comparison measured with AUC and EF values are depicted as Table 1.

Compared to WEGA, the screening performance of LSA were significantly improved. The mean AUC of DUD-E collection by LSA is 0.82, while WEGA gives a mean AUC of 0.74. LSA can achieve an average $EF^{1\%}$ of 27.0, which is about 30.4% higher than that of WEGA. We also calculated the median AUC. The median AUC of DUD-E collection by LSA is 0.84, while WEGA gives a median AUC of 0.72. All results of LSA and WEGA were treated with Wilcoxon signed rank test, $p < 0.001$. The virtual screening performances of 89.2% (91/102) libraries were improved with LSA, indicating that LSA had consistently better performance than WEGA. The detailed AUC results are provided in ESI (Table S1†). The ROC (receiver-operating characteristic) curves of the top-12 most performance improved targeted libraries (targeting UROK, CAH2, HDAC2, ADRB1, PYGM, ADRB2, MK10, PLK1, HIVINT, ACE, ROCK1 and THRB) virtual screenings using LSA and WEGA are depicted in Fig. 3, in which the curves in red are for LSA and the curves in black are for WEGA. The turning points of the curves are usually at the earlier stages of ROC curves, indicating that screening less than 20% of the compounds in a library can capture more than 80% intrinsic hits with LSA.

We further compared LSA with LS-align[21] and SPOT-ligand2 (ref. 22) which had been reported recently. It can be seen that LSA consistently had better performance as well. The $EF^{1\%}$ values by LSA are 22.7% and 12.0% higher than that by Flexi-LS-align and SPOT-ligand2 respectively. To further investigated the

Table 1 The virtual screening performances comparisons of WEGA, Rigid-LS-align, Flexi-LS-align, SPOT-ligand2 and LSA based on AUC and enrichment factors (EF) at top 1%, 5% and 10% of DUD-E

Method	AUC	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$
WEGA	0.74	20.7	7.5	4.4
Rigid-LS-align	—	20.1	6.9	4.3
Flexi-LS-align	0.75	22.0	7.2	4.5
SPOT-ligand2	—	24.1	8.6	5.2
LSA	**0.82**	**27.0**	**10.3**	**6.1**

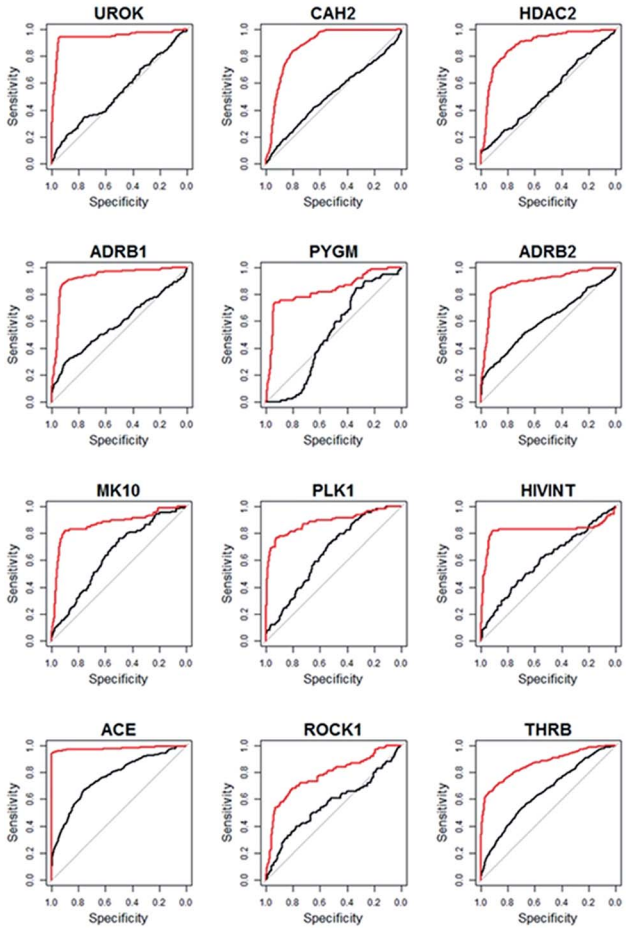

Fig. 3 The ROC curves of top-12 most performance improved targeted libraries virtual screenings using LSA and WEGA. The curves in red are for LSA and the curves in black are for WEGA.

§ The Directory of Useful Decoys (DUD-E) collection are available in the website, http://dude.docking.org/.

¶ LSA software and user guide can be downloaded for academic use at https://github.com/MingCPU/LSA.git

Table 2 EF values of WEGA, Rigid-LS-align and LSA on four protein categories of DUD-E

Categories (#proteins)	Method	$EF^{1\%}$	$EF^{5\%}$	$EF^{10\%}$
Kinases (26)	WEGA	17.7	6.4	3.8
	Rigid-LS-align	19	6.5	4.2
	LSA	**26.4**	**10.2**	**5.9**
Proteases (15)	WEGA	14.4	6.2	4.0
	Rigid-LS-align	15.4	6.3	4.3
	LSA	**24.9**	**11.3**	**6.5**
Nuclear receptors (11)	WEGA	**27.8**	**9.0**	5.4
	Rigid-LS-align	22.2	7.2	4.6
	LSA	22.3	8.9	**5.7**
GPCRs (5)	WEGA	9.6	3.8	2.7
	Rigid-LS-align	16.6	5.5	3.6
	LSA	**18.0**	**7.0**	**5.9**

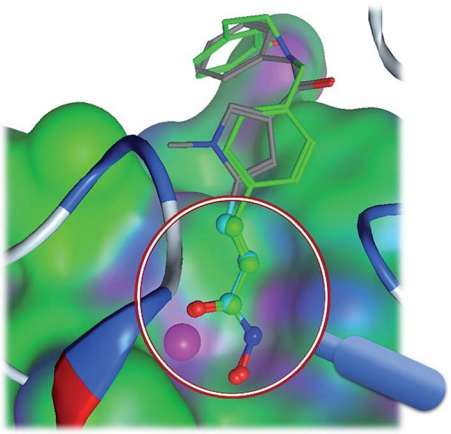

Fig. 4 The superimposed structures. The core substructures are superimposed in the magnifier. The molecule in green is CHEMBL343068 and the other molecule is CHEMBL275089 as in Fig. 1.

performance within DUD-E, we split DUD-E collection into four categories,[21] including kinases, proteases, nuclear receptors and GPCRs. The EF results of WEGA, Rigid-LS-align and LSA are as depicted in Table 2.

Superimposing two three-dimensional structures with LSA

LSA can be named as a 3D-substructure search engine, which superimposes two steric structures with substructure match restrictions. As shown in Fig. 4, LSA superimposes a compound against a template HDAC inhibitor (CHEMBL275089). In WEGA, two molecules are superimposed using the entire molecular mass center as the focused point. In LSA, however, two molecules are superimposed using the core substructure as the focused center. LSA starts from standard orientations and optimizes with four possible unique initial alignments.[4] The superimposing is optimized toward the large volume of core substructures base on the weight assignment. Therefore, LSA can be used as a better tool to dock a molecule into a binding pocket for a co-crystal complex if the native ligand and the privileged substructure(s) or "a warhead" of the ligand is known.

Conclusions

LSA reflects the fact that a privileged substructure is more important than the rest of the chemical structure in a query/template structure in virtual screening a compound library. After validating LSA with 102 targeted compound libraries, we have proved that the three-dimensional substructure search algorithm does result in improved virtual screening performance.

However, there might exist multiple privileged core substructures in a query structure. LSA cannot handle these cases. Although, these cases are rare.

Successfully applying LSA depends also on correctly specifying a core query substructure in a template structure. A larger core query substructure may result in no hits. A user should figure out the balance point of this technology. Our experience indicates that LSA is more suitable for screening bioactive compounds with a "warhead", or a covalent binding group.[23]

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work has funded in part of the science & technology program of Guangzhou (201604020109), the science & technology planning project of Guangdong Province (2016A020217005), GD Frontier & Key Techn. Innovation Program (2015B010109004), GD-NSF (2016A030310228), the National Key R&D Program of China (2017YFB02034043), Guangdong Provincial Key Lab of Construction Foundation (2011A060901014), Natural Science Foundation of China (U1611261, 61772566, 81773636) and the fundamental research funds for the central universities under grant 17LGJC23.

Notes and references

1 X. Yan, C. Liao, Z. Liu, A. T. Hagler, Q. Gu and J. Xu, *Curr. Drug Targets*, 2016, **17**, 1580–1585.
2 A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain and B. Kelley, *J. Med. Chem.*, 2010, **53**, 3862–3886.
3 J. A. Grant and B. T. Pickup, *J. Phys. Chem.*, 1995, **99**, 3503–3510.
4 X. Yan, J. B. Li, Z. H. Liu, M. H. Zheng, H. Ge and J. Xu, *J. Chem. Inf. Model.*, 2013, **53**, 1967–1978.
5 M. Sastry, J. F. Lowrie, S. L. Dixon and W. Sherman, *J. Chem. Inf. Model.*, 2010, **50**, 771–784.
6 D. H. S. Raymond, E. Carhart and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64–73.
7 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
8 J. Xu, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 25–38.
9 G. M. Sastry, S. L. Dixon and W. Sherman, *J. Chem. Inf. Model.*, 2011, **51**, 2455–2466.

10 H. Cai, T. Wang, Z. Yang, Z. Xu, G. Wang, H. Y. Wang, W. Zhu and K. Chen, *J. Chem. Inf. Model.*, 2017, **57**, 2329–2335.
11 O. Ivanciuc, *Curr. Comput.-Aided Drug Des.*, 2013, **9**, 153–163.
12 Y. Hu, D. Stumpfe and J. Bajorath, *F1000Research*, 2013, **2**, 199.
13 H. Peng, Z. Liu, X. Yan, J. Ren and J. Xu, *Sci. Rep.*, 2017, **7**, 11121.
14 S. Barelier and I. Krimm, *Curr. Opin. Chem. Biol.*, 2011, **15**, 469–474.
15 D. P. Dowling, S. L. Gantt, S. G. Gattis, C. A. Fierke and D. W. Christianson, *Biochemistry*, 2008, **47**, 13554–13563.
16 H. Park, S. Kim, Y. E. Kim and S. J. Lim, *ChemMedChem*, 2010, **5**, 591–597.
17 P. J. Ballester and W. G. Richards, *J. Comput. Chem.*, 2007, **28**, 1711–1723.
18 J. Kirchmair, S. Distinto, P. Markt, D. Schuster, G. M. Spitzer, K. R. Liedl and G. Wolber, *J. Chem. Inf. Model.*, 2009, **49**, 678–692.
19 M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2012, **55**, 6582–6594.
20 J. Li, T. Ehlers, J. Sutter, S. Varma-O'brien and J. Kirchmair, *J. Chem. Inf. Model.*, 2007, **47**, 1923–1932.
21 J. Hu, Z. Liu, D.-J. Yu, Y. Zhang and A. Valencia, *Bioinformatics*, 2018, **34**, 2209–2218.
22 T. Litfin, Y. Zhou and Y. Yang, *Bioinformatics*, 2017, **33**, 1238–1240.
23 J. Du, X. Yan, Z. Liu, L. Cui, P. Ding, X. Tan, X. Li, H. Zhou, Q. Gu and J. Xu, *Bioinformatics*, 2017, **33**, 1258–1260.