Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2008) 1276–1286

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor

L.Y. Han [a], X.H. Ma [a], H.H. Lin [a], J. Jia [a], F. Zhu [a], Y. Xue [c], Z.R. Li [c], Z.W. Cao [b], Z.L. Ji [d], Y.Z. Chen [a,b,*]

[a] *Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore*

[b] *Shanghai Center for Bioinformation Technology, Shanghai 201203, PR China*

[c] *College of Chemistry, Sichuan University, Chengdu 610064, PR China*

[d] *Bioinformatics Research Group, School of Life Sciences, Xiamen University, Xiamen 361005, FuJian Province, PR China*

Received 29 April 2007; received in revised form 5 December 2007; accepted 5 December 2007
Available online 15 December 2007

Abstract

Support vector machines (SVM) and other machine-learning (ML) methods have been explored as ligand-based virtual screening (VS) tools for facilitating lead discovery. While exhibiting good hit selection performance, in screening large compound libraries, these methods tend to produce lower hit-rate than those of the best performing VS tools, partly because their training-sets contain limited spectrum of inactive compounds. We tested whether the performance of SVM can be improved by using training-sets of diverse inactive compounds. In retrospective database screening of active compounds of single mechanism (HIV protease inhibitors, DHFR inhibitors, dopamine antagonists) and multiple mechanisms (CNS active agents) from large libraries of 2.986 million compounds, the yields, hit-rates, and enrichment factors of our SVM models are 52.4–78.0%, 4.7–73.8%, and 214–10,543, respectively, compared to those of 62–95%, 0.65–35%, and 20–1200 by structure-based VS and 55–81%, 0.2–0.7%, and 110–795 by other ligand-based VS tools in screening libraries of $\geq$1 million compounds. The hit-rates are comparable and the enrichment factors are substantially better than the best results of other VS tools. 24.3–87.6% of the predicted hits are outside the known hit families. SVM appears to be potentially useful for facilitating lead discovery in VS of large compound libraries.

Keywords: Computer aided dug design; Drug discovery; High-throughput screening; Lead discovery; Machine learning method; Virtual screening

1. Introduction

Virtual screening (VS) has been extensively explored for facilitating lead discovery [1–4] and for identifying agents of desirable pharmacokinetic and toxicological properties [5,6]. Machine learning (ML) methods have recently been used for developing ligand-based VS (LBVS) tools [7–14] to complement or to be combined with structure-based VS (SBVS) [1,15–26] and other LBVS [2,27–30] tools aimed at improving the coverage, performance and speed of VS tools.

ML methods have been used as part of the efforts to overcome several problems that have impeded progress in more extensive applications of SBVS and LBVS tools [1,31]. These problems include the vastness and sparse nature of chemical space needs to be searched, limited availability of target structures (only 15% of known proteins have known 3D structures), complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects. LBVS may in some cases limit the diversity of hits due to the bias of training molecules [15]. Therefore, alternative approaches that enhance screening speed and compound diversity without relying on target structural information are highly desired. ML methods have been explored for developing

* Corresponding author at: Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore. Tel.: +65 6874 6877; fax: +65 6774 6756.
E-mail address: phacyz@nus.edu.sg (Y.Z. Chen).

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.12.002

Table 1
Comparison of the reported performance of different virtual screening (VS) methods in screening large libraries of compounds

Type of VS method and size of compound libraries screened	VS method [references]	Compounds screened			Virtual hits selected by VS method		Known hits selected by VS method			
		No of compounds	No of known hits	Percent of known hits	No of compounds selected as virtual hits	Percent of screened compounds selected as virtual hits	No of known hits selected	Yield	Hit-rates	Enrichment factor
Structure-based VS, extremely large libraries (≥1 M)	Docking + pre-screening filter [2,18,19]	1 M–2 M	355–630	∼0.03%	1 K–60 K	0.08–3%	340–390	62–95%	0.65–35%	20–1200
Structure-based VS, large libraries	Docking + pre-screening filter [11,20–26]	134 K–400 K	100–1016	0.12–0.76%	375–4.5 K	0.28–3%	5–231	2–30%	0.11–17%	4–66
Ligand-based VS (machine learning), extremely large libraries (≥1 M)	Machine learning–SVM [2,8,11,13]	2.5 M	22–46	0.0009–0.0018%	2.5 K–11 K	0.1–0.45%	18–25	55–81%	0.2–0.7%	110–795
Ligand-based VS (machine learning), large libraries	Machine learning–SVM [2,9]	172 K	118–128	∼0.07%	1.7 K	1%	26–70	22–55%	1.5–4.1%	22–55
	Machine learning–SVM [11,12]	98.4 K	259–1146	0.26–1.16%	984	1%	131–710	44–69%	14–72%	44–69
	Machine learning–BKD [12,9,11,13,14]	101 K–103 K	259–1166	0.25–1.2%	5.1 K	5%	65–972	14–94%	1.2–18.9%	3–19
	Machine learning–LMNB [1,11,13]	172 K	118	0.069%	1.7 K	1%	19	16%	1%	15
	Machine learning–CKD [18,12]	98.4 K	259–1211	0.26–1.23%	984	1%	132–960	34–94%	13–98%	53–94
Ligand-based VS (clustering), large libraries	Hierachical k-means [5,28]	344.5 K	91–1556	0.026–0.45%	3750–21285	1.1–6.2%	27–761	23–55%	0.72–5%	7.97–31.2
	NIPALSTREE [5,28]	344.5 K	91–1556	0.026–0.45%	3469–28125	1.0–8.2%	17–625	18–50%	0.49–2.8%	3.51–18.7
	Hierachical k-means + NIPALSTREE disjunction [5,28]	344.5 K	91–1556	0.026–0.45%	7317–43165	2.1–12.3%	30–980	33–72%	0.41–2.9%	4.86–17.6
	Hierachical k-means + NIPALSTREE conjunction [5,28]	344.5 K	91–1556	0.026–0.45%	538–6692	0.16–1.9%	14–406	6–32%	1.1–10.2%	7.77–98
Ligand-based VS (structural signatures), extremely large libraries (≥1 M)	Pharmacophore [3,29,80,81]	1.77 M–3.8 M	55–144	0.0014–0.0081%	20 K–1 M	1.15–26%	6–39	11–70%	0.0039–0.084%	3–10.3
Ligand-based VS (structural signatures), large libraries	Pharmacophore [1,30]	380 K	30	0.0079%	6917	1.82%	23	76.7%	0.33	41.8