BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(11)

An alphabetic code based atomic level molecular similarity search in databases

Nallusamy Saranya & Samuel Selvaraj*

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirapalli – 620024, Tamilnadu, India; Samuel Selvaraj – Email: selvarajsamuel@gmail.com; Phone: +91-431-2407071; Fax: +91-431-2407045; *Corresponding author

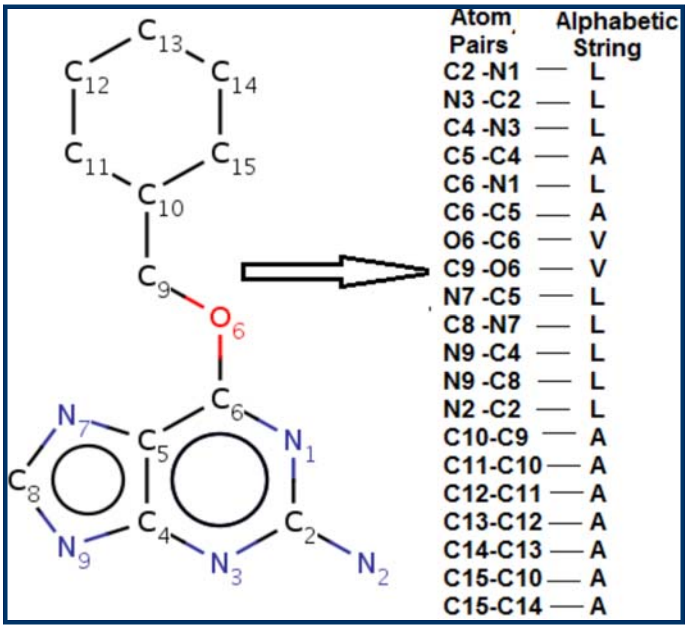
Received May 14, 2012; Accepted May 27, 2012; Published June 16, 2012

Abstract:
Atomic level molecular similarity and diversity studies have gained considerable importance through their wide application in Bioinformatics and Chemo-informatics for drug design. The availability of large volumes of data on chemical compounds requires new methodologies for efficient and effective searching of its archives in less time with optimal computational power. We describe an alphabetic algorithm for similarity searching based on atom-atom bonding preference for ligands. We represented 170 cyclin-dependent kinase 2 inhibitors using strings of pre-defined alphabets for searching using known protein sequence alignment tools. Thus, a common pattern was extracted using this set of compounds for database searching to retrieve similar active compounds. Area under the receiver operating characteristic (ROC) curve was used for the discrimination of similar and dissimilar compounds in the databases. An average retrieval rate of about 60% is obtained in cross-validation using the home-grown dataset and the directory of useful decoys (DUD, formally known as the ZINC database) data. This will help in the effective retrieval of similar compounds using database search.

Key words: Atom pair, CDK-2, similarity searching, molecular similarity

Background:

Molecular similarity and diversity studies have gained importance through their wide application in the field of bio-informatics and chemo-informatics **[1, 2]**. The main goal of structure-based drug design (SBDD) is to find novel lead compounds with potent and specific activity. Based on the principle "similar molecules exert similar activity", ligand similarity searching has gained importance in virtual screening strategy **[3, 4]**. Ligand similarity can be assessed by means of comparing their structures using 1D, 2D and 3D approaches such as tanimoto coefficient **[2, 5]**, SMILES **[6]**, COMFA **[7]**, COMSIA **[8]** etc **[1, 9, 10]**. While, 1D descriptors explain the chemical nomenclature, physicochemical and biological properties, 2D descriptors provide information regarding the fragment counts, topological indices, molecular connectivity and graphical representation and 3D descriptors detail molecular surface, volume and interaction energies. Each descriptor has its own importance in the search of related molecules. Large numbers of descriptors have been reported to date and software are available for the calculation of descriptors (CODESSA **[11]**, DRAGON **[12]**, Molinspiration **[13]** and COMFA **[7]** etc). Atom pair descriptors and topological descriptors are very popular in 2D similarity searching for more than past two decades **[14-16]**. Atom pair descriptors encode all atoms in a molecule together with the length of the shortest bond-by-bond path between them. Topological descriptors are single valued descriptors that can be calculated from the 2D graph representation of molecule **[15, 17]**. Earlier Bremser **[18]** has described an encoding system HOSE (Hierarchical Ordered description of the Substructure Environment) code for NMR spectra prediction. This system describes the structural neighbors of the particular atom of interest, which in NMR essentially identifies those atoms within the molecule influencing the chemical shift of that atom. Grant et al **[19]** have implemented Lingos approach to measure molecular similarity by converting the molecule into a set of strings. In this approach molecular pair similarity is assessed based on the occurrence of substring frequency. In addition to similarity searching, atom type and bond type information plays an important role in molecular mechanics calculations, QSAR and QSPR studies **[20, 21]**. A recent study has introduced a new type of atom pair descriptor namely bond pair descriptor which includes element type, hybridization state, aliphatic/aromatic character, and cyclic/acyclic arrangement information for ligand similarity searching **[22].** Extended-connectivity fingerprints method uses circular fingerprints for representing molecular features relevant to molecular activity which can be used for clustering, similarity searching and virtual screening **[23]**.



Atom Pairs		Alphabetic String
C2 -N1	—	L
N3 -C2	—	L
C4 -N3	—	L
C5 -C4	—	A
C6 -N1	—	L
C6 -C5	—	A
O6 -C6	—	V
C9 -O6	—	V
N7 -C5	—	L
C8 -N7	—	L
N9 -C4	—	L
N9 -C8	—	L
N2 -C2	—	L
C10-C9	—	A
C11-C10	—	A
C12-C11	—	A
C13-C12	—	A
C14-C13	—	A
C15-C10	—	A
C15-C14	—	A

Figure 1: Alphabetic representation of atoms pairs in (PDBID_HTMID) 1E1V_CMG.

Ligand similarity is also measured based on the spatial alignment with atomic property fields and (a generalized 3D pharmacophoric potential) was tested on a large diverse dataset of 115 protein-ligand complexes **[24].** Verma et al **[25]** have developed quantitative structure activity relationship model based on local similarity indices to understand the binding mechanism and to improve biological activity. Algorithms related to finding molecular matched pairs, where pair of compounds differs only in single localized structural change have been analyzed in the large volume chemical dataset **[26]**. Chemical similarity has also been analyzed based on the fragment profiles in class specific and class independent compounds which produced better results in database screening **[27].** Existing methodologies vary in their performance for different targets and also select varied set of actives for specific target. Accumulation of huge chemical compounds in the databases necessitates the development of new similarity methods in finding actives for the particular target protein **[28, 29]**. Mostly, biological molecules such as proteins, nucleic acids and small molecular ligands are mainly made up of fundamental elements like carbon (C), Hydrogen (H), nitrogen (N) and oxygen (O). The representation of amino acid sequence information in terms of 20 simple letter alphabetic codes and 4 codes for nucleic acids has provided a good solution to the efficient storage and retrieval of molecular sequence data. Also the development of powerful algorithms and the widely used tool for performing sequence analysis such as BLAST (Basic Local Alignment Search Tool) **[30]** to mine biologically related sequences very quickly and efficiently has been at the core of bioinformatics analysis of genome as well protein sequences that amount to millions of character strings. BLAST remains the fundamental resource for most of the bioinformatics approaches like gene prediction, structure prediction, function annotation etc. There is no way of representing molecular structures of ligands in terms of an alphabetic code. In the present work, we have developed an amino acid-like alphabetic code to represent atom-atom bonding preference in ligands to search for similar ligand molecules in databases. Bonds between atoms are important and remain the fundamental characteristics of similar molecules. BLAST program was used for the search and retrieval of similar molecules. We have implemented our method for Cyclin-dependent kinase-2 (CDK-2) inhibitors. CDK-2 is one of the active targets in SBDD and is involved in regulation of cell cycle proliferation and RNA polymerase II (RNAP II) transcription cycle **[31].** Earlier studies related to pharmocophore development have been reported for CDK-2 inhibitors **[32-35]** Key features such as hydrogen bond donors, acceptors and hydrophobic feature required for the activity have been reported and in addition steric effects and docking were performed to enhance the retrieval rate of active compounds. Alphabetic representation of atom-atom bonding in ligands (inhibitors) provides an easy way of analysis using sequence alignment tools for similarity searching of CDK-2 inhibitors. A consensus pattern was derived from the alignment results and this pattern was used as query for database searching. We use BLAST tool to perform ligand similarity search to retrieve actives similar to query pattern from the database. To test the efficiency of approach, similarity searching has been performed on the compounds having diverse scaffolds and similar scaffolds. Statistical validation of the method was performed by database searching using area under the receiver operating characteristic (ROC) curve **[36]**.

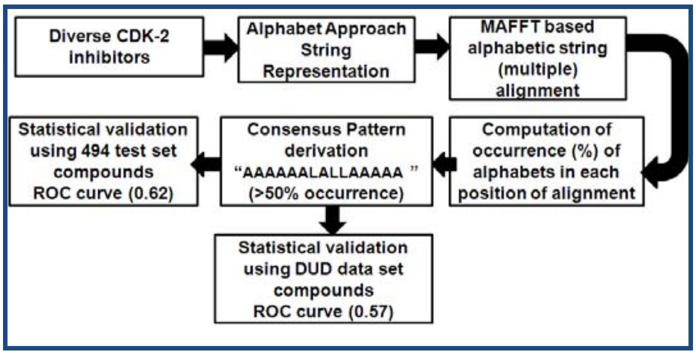

Diverse CDK-2 inhibitors
Alphabet Approach String Representation
MAFFT based alphabetic string (multiple) alignment
Computation of occurrence (%) of alphabets in each position of alignment
Consensus Pattern derivation "AAAAAALALLAAAAA " (>50% occurrence)
Statistical validation using 494 test set compounds ROC curve (0.62)
Statistical validation using DUD data set compounds ROC curve (0.57)

Figure 2: Workflow of the alphabet algorithm

Methodology:

Alphabet Representation

In the present work, single letter alphabetic codes based on atom-atom bonding (atom pairs) pairs in ligand structures has been assigned, which corresponds to particular amino acid in protein sequences. For example, when a carbon atom is attached to another carbon (C-C) atom, it is represented as "A", like wise when oxygen bonded with nitrogen or vice versa (O-N or N-O), it is represented as Q. If any halogen atom is bonded with C atom, it is represented as alphabet "M". Alphabets are assigned in such a way that atom pairs involving carbon atom are given codes of non-polar aliphatic amino acids, atom pairs involving nitrogen atom are provided with polar uncharged amino acids, oxygen atom with aromatic amino acids and hydrogen atom with positively charged amino acids such that scoring is appropriate for alignment. The alphabetic assignment for different atom-atom pairs in a ligand has been given in **Table 1 (See supplementary material).** As an example, the alphabetic representation for ligand SBC (PDB ID: 2BKZ) is given in (**Figure 1**). As alphabets assigned are in convention with the property of amino acids, BLOSUM 62 **[37]** scoring matrix used in sequence alignment tools has been applied to score and retrieve similar alphabetic strings. 170 CDK-2 inhibitors (reported in our earlier study) **[38]** obtained from Protein Data Bank were converted to alphabet strings representation using an in house perl program. After removal of the ATP molecules and redundant inhibitors, these strings were aligned using MAFFT **[39]** sequence alignment program. As gapped alignment is performed with the alphabet strings, highly similar substrings are aligned with high score irrespective of the order of the alphabets. From the alignment a consensus feature was derived based on the position of each alphabet in the string. Inhibitors that lead to the improper alignment of alphabetic strings were removed. Finally, 138 CDK-2 inhibitors were used for the common feature derivation. An alphabet which occurs more than 50% in each position of the multiple sequence alignment was taken as a threshold to derive consensus pattern (common feature). Consensus pattern obtained was used as a query to search databases for actives and inactives.

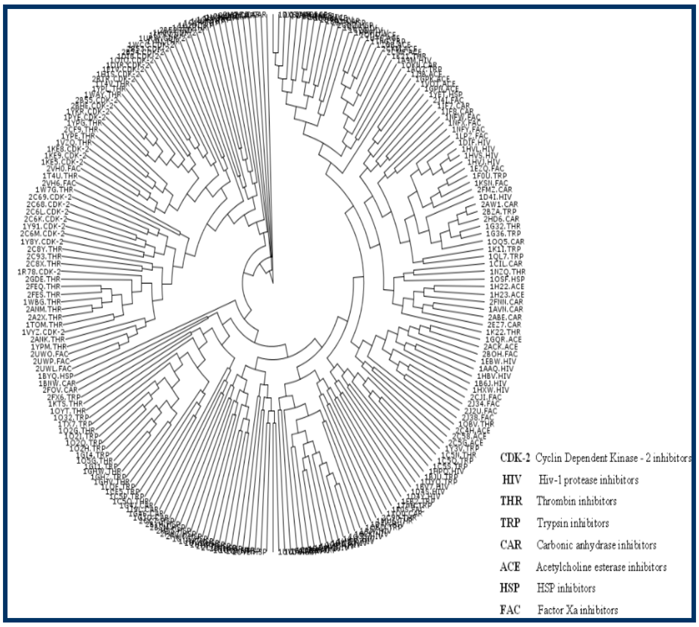
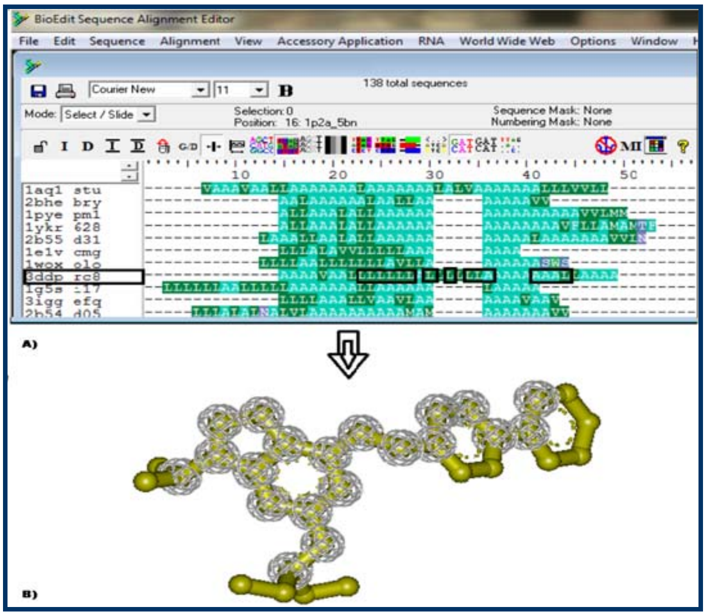

CDK-2 Cyclin Dependent Kinase - 2 inhibitors
HIV Hiv-1 protease inhibitors
THR Thrombin inhibitors
TRP Trypsin inhibitors
CAR Carbonic anhydrase inhibitors
ACE Acetylcholine esterase inhibitors
HSP HSP inhibitors
FAC Factor Xa inhibitors

Figure 3: Clustering of alphabetic strings of 200 ligands

Database searching and Statistical validation

Validation of the methodology was performed using two databases for the search of CDK-2 actives. In the earlier study, multi-complex (protein-ligand complex) based and most-frequent-feature pharmocophore map model was validated using 494 compounds dataset which includes 194 active inhibitors and 300 non- inhibitors for the target CDK-2. The 494 compounds dataset **[34]** and Directory of Useful decoys (DUD, formally known as the ZINC database) **[40]** dataset reported for CDK-2 were used separately as a test set for validating the present alphabet approach. DUD data set contains 72 actives and for each active 36 decoys with similar physical properties (e.g. molecular weight, calculated LogP) but dissimilar topology was reported. These compounds in these databases were also represented as strings and the consensus pattern extracted was used as the query to perform database search using BLAST module in Bioedit **[41]** software. Statistical validation was performed using receiver operating characteristic (ROC) curve using SPSS 16.0 software and its significance is assessed by the computation of P value under the null hypothesis that the area under the curve equals 0.5. The flow chart in **(Figure 2)** provides a brief overview of the present work.


BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window
Courier New
11
B
Mode: Select / Slide
A)
B)

138 total sequences

Selection: 0
Position: 16: 1p2a_5bn

Sequence Mask: None
Numbering Mask: None

```
                  10        20        30        40        50
1aq1 stu  ------VAAAVAALLAAAAAAAALAAAAAAAALALVAAAAAAAALLLVVLL------
2bhe bry  ----------------AALAAAAAAALAALLAA-----AAAAAVV----------
1pye pml  ----------------ALLAAALALLAAAAAA-----AAAAAAAAAAAVVLMM------
1ykr 628  ----------------ALLAAALALLAAAAAA-----AAAAAAAAAVFLLAMAMTF----
2b55 d31  --------------LAAALLAALALLAAAAAA-----AAAAALAAAAAAAAVVLN----
1e1v cmg  ----------------LLLALAVVLLLLLAAA-----AAAA-----------
1wox olo  --------------LLLLAALLLLLLLAVLLA-----AAAAAAASWS--------
3ddp rc8  ----------------AAAAVAALLLLLLLLLLLLLLLLAAAAAAALLAAAA------
1g5s i17  --LLLLLLAALLLLLLAAAAAAAALLAAAAAA-----LAAAAA---------
3igg efq  ----------------LLLLAAALLVAAVLAA-----AAAAVAAV-------
2b54 d05  -----LLLALALNALVLAAAAAAAAAAAMAM-----AAAAAAAVV------
```

Figure 4: (A) Common features marked in ligand RC8 has been boxed with dark line in the strings alignment; **(B)** Common features mapped (marked as spheres) in the ligand RC8 structure.

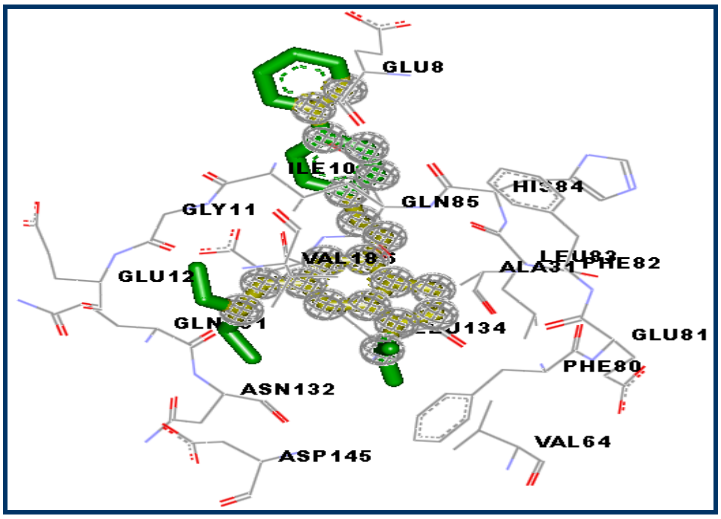


Figure 5: Common features (marked as spheres) of ligand RC8 reported in the (PDBID) 3DDP binding site. Nitrogen atom marked in yellow favors hydrogen bonding.

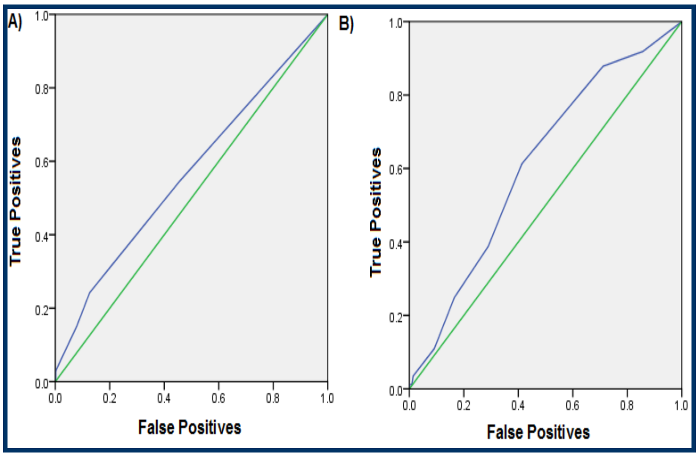


Figure 6: ROC curve for the retrieval of actives in **(A**) DUD and **(B)** Local database

Discussion:

At the initial stage of alphabet algorithm development, we have generated strings for each of the 200 ligand structures belonging to eight different protein families (data set from reference **[42]**). We employed the sequence alignment method to find how these strings are aligned and to explore how related ligands are grouped on alignment with different ligands. MAFFT based sequence alignment was performed and these alignments were clustered to find how these strings are grouped. In the cluster result we observed groupings among the similar ligands. Cluster diagram of the 200 ligands is given in (**Figure 3**). In a similar way, we tested the present approach by taking in to account non-bonded atoms in which an alphabet is assigned in such a way that nearest one atom is left and the next atom is taken in to consideration. Likewise, nearest two atoms were left and the next atom is taken in to account in other case. In the above mentioned two cases, no reasonable similarity was observed among related ligands. Hence, we considered atom-atom bonding (atom pairs) to do further analysis. With the above interesting observation, we applied alphabet approach for similarity searching using common feature obtained from CDK-2 inhibitors. 138 CDK-2 inhibitors with diverse scaffold belonging to different chemical classes were used for the consensus feature derivation. As mentioned afore, alignment of 138 CDK-2 inhibitors for obtaining common feature reported the pattern "AAAAAALALLAAAAA". This common feature has been marked in the ligand RC8 with respect to its position in the multiple sequence alignment of 138 CDK-2 inhibitors **(Figure 4).** Each alphabet in the conserved feature string reports the atom-atom bonding preferred in the whole CDK-2 inhibitors. Conserved features of ligand RC8 have been marked as spheres in (**Figure 4**). The reported pattern also provides relevant information regarding intermolecular interactions. For example, common feature pattern having the alphabet "A" represents the C-C bonding in the ligand which has the possibility to form hydrophobic interaction with the protein. Alphabet "L" represents the C-N bonding in the ligand which has possibility to form hydrogen bond (hydrogen donor or acceptor) with the protein. Residue interactions possible with the common feature were analyzed in (PDBID_HTMID) 3DDP_RC8 crystal structure and have been mapped in the CDK-2 binding site **(Figure 5).** This occupied region favors the highly conserved interactions with residues such as ALA 31, LEU 83 and LEU 134 which has been reported earlier **[38]**. Protein BLAST (Bioedit) was performed using the reported query "AAAAAALALLAAAAA" to search for occurrence of similar alphabetic (bonding pattern) strings in the databases. As gapped alignment is possible between query and matched alphabetic string from database, there exists a high chance to score and retrieve similar CDK-2 actives. ROC area under the curve (AUC) value of 0.57 and 0.62 was obtained on validation with DUD data set and 494 data set compounds respectively. **(Figure 6)** provides the details of retrieval rate of actives (true positives) and inactives (false positives) in the databases. P value was observed to be highly significant for the 494 data set (0.0) compared to DUD data set (0.20).

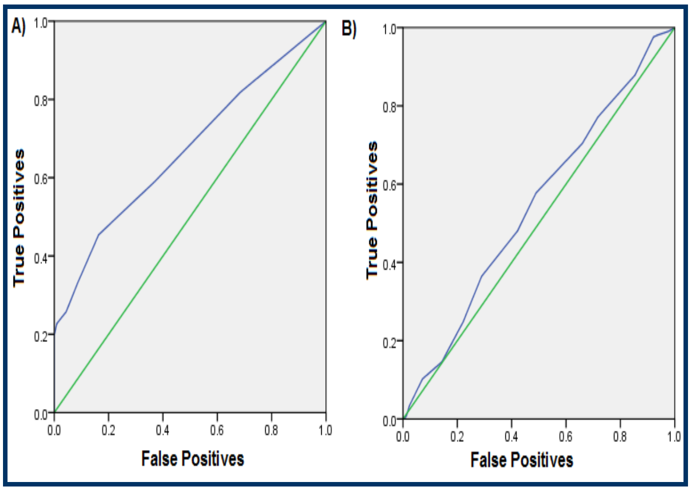


Figure 7: ROC curve for the retrieval of actives using query developed from 13 compounds in **(A)** DUD database and **(B)** Local database

In addition to derivation of common feature from diverse CDK-2 inhibitors, we have also implemented our alphabetic approach in 13 CDK-2 inhibitors **[43]** with similar scaffold obtained from The Binding database (Binding DB) **[44]** with 0.7 tanimoto similarity. As mentioned afore, alignment was performed to derive the consensus feature "AALLLLLAVAAAAAAALLAAAALLTFFVVM" and database searching was performed on the local and DUD databases. In this protocol, both the DUD and local databases were added with 13 compounds which are used in the consensus feature development. BLAST results reported the 13 compounds at the top hit list with high score compared to other actives in both databases. Area under ROC values of 0.67 and 0.55 **(Figure 7)** with the curve above the reference line (0.5) and P-value of 0.00 and 0.09 for the DUD and local databases were obtained respectively. In this model, though common feature were derived from the similar compounds, database searching has resulted in the retrieval of actives with diverse structure. In summary, the present alphabetic approach reported most of the important features required for intermolecular interaction in 138 CDK-2 inhibitors and hence was able to retrieve 60% (app.) of active compounds in database virtual screening. Common features of similar scaffold compounds reported maximum features for the activity such that it performed better in database searching compared to the common feature of diverse scaffold compounds. Incorporation of additional atom pair descriptors in alphabetic form (such as atom type, bond type, accessibility etc) with simple metrics will be performed in future algorithm development.

Conclusion:

We described the use of an alphabetic approach depending on the type of atom-atom pairs in a molecule for the purpose of representation of molecular structures for molecular similarity search in huge databases. This approach finds application in the initial screening of huge databases with computational time and complexity. It should be noted that bonding connectivity information and protein sequence matrices for alignment were not included in its current form. The present approach will be modified further by incorporating scoring matrix to retrieve hits with improved accuracy rate.

Acknowledgement:

We thank University Grants Commission, New Delhi for the award of RGNFS fellowship to N.S.

References:

[1] Maldonado AG *et al. Mol Divers.* 2006 **10**: 39 [PMID: 16404528].

[2] Willett P *et al. J Chem Inf Comput Sci.* 1998 **38**: 983

[3] Kubinyi H, *J Braz Chem Soc.* 2002 **13**: 717

[4] Kubinyi H, *Perspect Drug Discovery Des.* 1998 **11**: 225.

[5] Flower DR, *J Chem Inf Comput Sci.* 1998 **38:** 379

[6] Weininger D, *J Chem Inf Comput Sci.* 1988 **28**: 31.

[7] Cramer RD *et al. J Am Chem Soc.* 1988 **110**: 5959 [PMID: 22148765]

[8] Good AC *et al. J Med Chem.* 1993 **36**: 433 [PMID: 8474098].

[9] Nalewajski RF & Parr RG, *Proc Natl Acad Sci USA.* 2000 **97**: 8879 [PMID: 10922049].

[10] Kearsley SK *et al. J Chem Inf Comput Sci.* **36**: 118.

[11] Katritzky AR *et al.* CODESSA Reference Manual Version 20, University of Florida 1996.

[12] http://www.disat.unimib.it/chm/Dragon.htm

[13] www.molinspiration.com

[14] Carhart RE *et al. J Chem Inf Comput Sci.* 1985 **25**: 64.

[15] Kier LB & Hall LH, *J Chem Inf Comput Sci.* 2000 **40**: 792 [PMID: 10850784]

[16] Bender A *et al. J Chem Inf Comput Sci.* 2004 **44**: 1708 [PMID: 15446830]

[17] Randic M, *J Am Chem Soc.* 1975 **97**: 6609.

[18] Bremser W, *Anal Chim Act.* 1978 **103**: 355.

[19] Grant JA *et al. J Chem Inf Model.*2006 **46**: 1912 [PMID: 16995721]

[20] Wang J *et al.* J *Mol Graphics Model.* 2006 **25**: 247 [PMID: 16458552]

[21] Kier LB *et al. J Med Chem.* 1975 **18**:1272 [PMID: 1238571]

[22] Ahmed HE *et al. J Chem Inf Model.* 2010 **50**: 487 [PMID: 20232887]

[23] Rogers D & Hahn M, *J Chem Inf Model.* 2010 **50**: 742 [PMID: 20426451]

[24] Grigoryan AV *et al. J Comput Aided Mol Des***.** 2010 **24**: 173 [PMID: 20229197]

[25] Verma J *et al. J Chem Inf Model.* 2009 **49**: 2695 [PMID: 19994892]

[26] Hussain J & Rea C, *J Chem Inf Model.* 2010 **50**: 339 [PMID: 20121045]

[27] Batista J & Bajorath J, *J Chem Inf Model.* 2007 47:59 [PMID: 17238249]

[28] Sheridan RP & Kearsley SK, *Drug Discov Today.* 2002 **7**: 903 [PMID: 12546933]

[29] Zavodszky MI, *et al. J Mol Recognit.* 2009 **22**: 280 [PMID: 19235177]

[30] Sausville EA, *Trends Mol Med.* 2002 **8**:S32 [PMID: 11927285]

[31] Hecker EA, *et al. J Chem Inf Comput Sci.* 2002 **42**:1204 [PMID: 12377010]

[32] Toba S *et al. J Chem Inf Model.* 2006 46:728 [PMID: 16563003]

[33] Vadivelan S *et al. J Chem INF Model.* 2007 **47**:1526 [PMID: 17523616].

[34] Zou J *et al. J Mol Graph Model.* 2008 **27**: 430 [PMID: 18786843]

[35] Altschul SF *et al. J Mol Biol.* 1990 **215**: 403 [PMID: 2231712]

[36] Bamber D, *Journal of Math Psychol.* 1975 **12**: 387

[37] Henikoff S & Henikoff JG, *Proc Natl Acad Sci USA.* 1992 **89**: 10915 [PMID: 1438297]

[38] Saranya N & Selvaraj S, *Chem biol drug des.* 2011 **78**: 361 [PMID: 21599856]

[39] Katoh K et al *Nucleic Acids Res.* 2002 **30**: 3059 [PMID: 12136088]

[40] Huang N *et al. J Med Chem.* 2006 **49**: 6789 [PMID: 17154509]

[41] Hall TA, *Nucl Acids Symp Ser.* 1999 **41**: 95.

[42] Saranya N & Selvaraj S, *Bioorg Med Chem Lett. 2009* **19**: 5769 [PMID: 19706358]

[43] Chu XJ *et al. J Med Chem.* 2006 **49**: 6549 [PMID: 17064073]

[44] Liu T *et al. Nucleic Acids Res.* 2007:**35**: D198 [PMID: 17145705]

Edited by P Kangueane

Citation: **Saranya & Selvaraj,** Bioinformation 8(11): 498-503 (2012)

Supplementary material:

Table 1: Alphabetic code assignment for different atom pairs in ligands

Atom	Alphabet
C-C	A
H-C or C-H	I
O-C or C-O	V
N-C or C-N	L
H-H	K
O-H or H-O	W
N-H or N-H	S
O-O	Y
N-O O-N	Q
N-N	N
(BR, I, CL, S, F) – C	M
(BR, I, CL, S, F) – H	R
(BR, I, CL, S, F) – O	F
(BR, I, CL, S, F) –N	T