Pacific Symposium on Biocomputing 12:133-144(2007)

COMPARATIVE QSAR ANALYSIS OF BACTERIAL-, FUNGAL-, PLANT- AND HUMAN METABOLITES.

EMRE KARAKOC

School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.

S. CENK SAHINALP

School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.

ARTEM CHERKASOV

Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, 2733, Heather street, Vancouver, BC, V5Z 3J5, Canada.

Several QSAR models have been developed using a linear optimization approach that enabled distinguishing metabolic substances isolated from human-, bacterial-, plant- and fungal- cells. Seven binary classifiers based on a k-Nearest Neighbors method have been created using a variety of 'inductive' and traditional QSAR descriptors that allowed up to 95% accurate recognition of the studied groups of chemical substances.

The conducted comparative QSAR analysis based on the above mentioned linear optimization approach helped to identify the extent of overlaps between the groups of compounds, such as cross-recognition of fungal and bacterial metabolites and association between fungal and plant substances. Human metabolites exhibited very different QSAR behavior in chemical space and demonstrated no significant overlap with bacterial-, fungal-, and plant-derived molecules.

When the developed QSAR models were applied to collections of conventional human therapeutics and antimicrobials, it was observed that the first group of substances demonstrate the strongest association with human metabolites, while the second group exhibit tendency of 'bacterial metabolite – like' behavior. We speculate that the established 'drugs - human metabolites' and 'antimicrobials – bacterial metabolites' associations result from strict bioavailability requirements imposed on conventional therapeutic substances, which further support their metabolite-like properties.

It is anticipated that the study may bring additional insight into QSAR determinants for human-, bacterial-, fungal- and plant metabolites and may help rationalizing design and discovery of novel bioactive substances with improved, metabolite-like properties.

1. Introduction

In a series of previous works we reported the use of our own 'inductive' and conventional 2D and 3D QSAR descriptors for creating binary QSAR models capable of recognizing various groups of substances including antimicrobial

molecules and peptides [1,2], steroid-like compounds [3], human therapeutics, drug-like chemicals [4] as well as bacterial and human metabolites [4,5]. These binary QSAR classifiers allowed defining certain structural determinants of the studied groups and provided important insights into their positioning in chemical space. Thus, the developed QSAR models could demonstrate immanent similarity between conventional antimicrobials and native bacterial metabolites and have been suggested as prospective tools for 'in silico' antibiotic discovery.

In the current study we applied similar QSAR approach to the broader spectra of bacterial-, human-, plant- and fungal metabolites that have been explored for mutual overlaps as well as for possible associations with classes of conventional human therapeutics, antimicrobials and biologically neutral drug-like chemicals.

2. Materials and Methods

2.1. *Molecular Datasets*

The dataset of antimicrobial compounds has been assembled from several public resources including *ChemIDPlus* service [6], *the Journal of Antibiotics* database [7] and from the literature [8-10]. The conventional drug molecules covering a broad range of therapeutic activities have all been identified from the *Merck Index Database* [11].

The structures of bacterial-, plant- and fungal metabolites have been obtained from the *AnalytiCon-Discovery* company [12]. Drug-like substances used in the study have been selected from the *Assinex Gold* collection [13]. Structures of human metabolites have been obtained from the *Metabolomics* database [14].

The redundancy of the resulting dataset containing 519 Antimicrobials, 958 Drugs, 1202 Drug-like substances with no known therapeutic effects, together with 1102 Human-, 551 Bacterial-, 2351 Plant- and 825 Fungal metabolites has been ensured through the SMILES records and by descriptors-based clustering. All molecular structures have been further optimized with the MMFF94 force-field [15] and using MOE modeling package [16].

2.2. *QSAR Descriptors*

The optimized structures of 7508 compounds have been used for calculating 26 non cross-correlating 'inductive' QSAR descriptors [1-5] and 33 conventional QSAR parameters (the corresponding descriptions can be found in Appendix).

The resulting set of 59 QSAR parameters descriptors computed for 7508

studied compounds has been normalized for [0,1] range. The normalized values have then been used to generate QSAR models distinguishing all four types of natural metabolites under study.

2.3. *Mathematical Approach*

For the purpose of distinguishing four types of the studied metabolite substances based on descriptors we have utilized the *k*-Nearest Neighbors (*k*-NN) classification approach. This method requires the definition of distance *D(S, R)* between any pair of molecules S and R in *d*-dimensional descriptors space. According to QSAR formalism, such a distance measure should reflect functional association and/or chemical similarity between the molecules. Thus, the *k*-NN approach allows descriptors-based clustering of chemical compounds according to already known biological activity and can be used to classify an untested chemical substance by its proximity to established clusters. Given a distance function *D( )*, searching for the *k*-Nearest Neighbors of untested chemical substance requires comparing its distance with tested compounds which is computationally costly. The efficiency of the classification process can be improved by efficient data-structures developed for metric spaces. Thus, a metric distance function is used for our clustering of tested compounds. A distance measure *D( )* forms a metric if the following conditions are satisfied.

D(S, S) = 0: a point has distance 0 to itself.
D(S, R) = D(R, S): distance is symmetric.
D(S, R) $\leq$ *D(S, Q) + D (Q, R):* distance satisfies the triangle inequality.

The distance measures satisfying the above conditions include Hamming Distance (i.e. L_1): $\sum_{h=1}^{d}|S_h - R_h|$, Euclidean Distance (i.e. L_2): $\sqrt{\sum_{h-i}^{d}(S_h - R_h)^2}$, Maximum of dimensions (i.e. L_∞): $\max_{1\leq h\leq d}|S_h - R_h|$ among others.

In the current work we utilized the weighted Hamming Distance $\sum_{h=1}^{d}\sigma_h|S_h - R_h|$, ($\forall \sigma_i, \sigma_i \geq 0$) that allows differentiating relevance of various QSAR descriptors for a given activity. Weighted Hamming Distance representations allows establishing optimal σ_i values that maximize separation of active $T^A = \{T^A_1, T^A_2 \ldots T^A_m\}$ and inactive $T^I = \{T^I_1, T^I_2 \ldots T^I_n\}$ elements in the training set: $T = T^A \bigcup T^I$.
We utilized the Linear Programming approach to minimize the function,

$$f(T) = \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{h=1}^{d}\sigma_h \cdot \left|T_i^A[h] - T_j^A[h]\right|\right)/m^2 + \left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{h=1}^{d}\sigma_h \cdot \left|T_i^I[h] - T_j^I[h]\right|\right)/n^2 - \left(\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{h=1}^{d}\sigma_h \cdot \left|T_i^A[h] - T_j^I[h]\right|\right)/m \cdot n$$

such that the following three conditions are satisfied:

$$\forall T_i^A \in T^A \left(\sum_{j=1}^{m}\sum_{d=1}^{h}\sigma_h \cdot \left|T_i^A[h] - T_j^A[h]\right|\right)/m^2 \le \left(\sum_{j=1}^{n}\sum_{d=1}^{h}\sigma_h \cdot \left|T_i^A[h] - T_j^I[h]\right|\right)/m \cdot n$$

$$\forall i \quad 0 \le \sigma_i \le 1$$

$\sum_{i=1}^{d}\sigma_i \le C$, where C - is a used-defined constant.

The aim of the clustering is determining best descriptor space where the average distance among compounds reflexes the functional similarity. Although the sensitivity and specificity can be improved using more restricting constraints, the optimization may end up with an infeasible or over-trained solution. In order to avoid infeasible solutions and overtraining, the average distance constraints are used.

Another important factor for the clustering is the size of training data. The accuracy of the clustering is going to improve logarithmically with the increasing size of training data. According to our observations, the ideal training dataset is 90% of the whole dataset.

More details on the adopted *k*-NN procedure can be found in [5, 17]. It should be outlined that the described mathematical procedure not only maximizes the average distance between active and inactive elements of the training set, but also aims to minimize the average within-the-class distance and, therefore, tends to condense activity-clusters.

3. Results and Discussion

The defined clusters of chemical compounds of interest in *d*-dimensions can then be used to characterize unknown entries (molecules) by projecting their QSAR parameters into descriptors space (Figure 1).

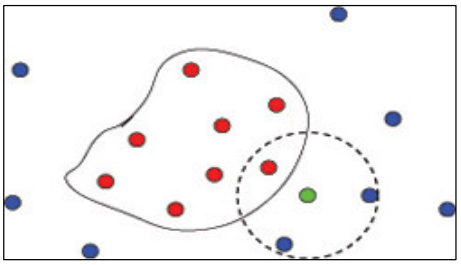

Figure 1. Projection of unknown compound (green point) onto chemical space where active compounds (red points) have been separated from inactive ones (blue) using *k*-NN algorithm.

In particular, an untested compound (green point on Figure 1) can be associated to a certain pre-defined activity cluster by considering affiliations of its *k*-nearest neighbors. In the current study we considered assigning the tested compound to the cluster of its nearest neighbor.

The linear optimization instance for determining the distance function is obtained from the input compound dataset as described above and represented using the MPS format. This linear programming instance is solved using the open-source linear programming solver CPLEX [18]. The data structure for searching the nearest neighbor of a query point was SC-Vantage Point Tree that we developed earlier [19]. All programs are implemented using the standard C/C++ libraries on UNIX environment.

We tested applicability of the above-described approach for creating binary QSAR classifiers that operate by 59 'inductive' and conventional QSAR variables. The combined molecular dataset consisting of 1202 drug-like chemicals, 958 conventional drugs of various types, 519 specific antimicrobials, as well as 551 Bacterial-, 2351 Plant-, 825 Fungal- and 1102 Human Metabolites (with 59 normalized QSAR descriptors assigned to each entry) has been used to create *k*-NN based QSAR models. Although 7 new QSAR models are developed (one for each type of chemical compounds) based on our combined dataset, we only concentrate on the following categories.

3.1. *Plant Metabolites*

To create a binary QSAR model accurately distinguishing plant metabolites from other types of chemical substances, we considered a set of 2351 natural compounds characterized from plant isolates by *AnalytiCon-Discovery* Company [12]. As the negative control for such model we considered a combination 1477 conventional human therapeutic substances (including 519 antibacterials), 1202 biologically inactive chemicals (that nonetheless justify the 'Lipinski's' drug-likeness rule) and 2478 metabolic substances participating in human, bacterial and fungal biological pathways. We included drugs and drug-like molecules into negative control to ensure that the desired QSAR model for plant-metabolites won't be simply biased toward drug-like structures. On another hand, the presence of other types of native metabolites in a negative control aimed to ensure that the QSAR approach won't be generally recognizing any metabolic substances.

To develop the *k*-NN based QSAR binary model (yes/no) for plant metabolites we assigned a bioactivity value of 1.0 (dependent variable) to 2351 plant substances and treated them as actives. All the remaining 5157 molecules

have been assigned null activities as *k*-NN algorithm attempted separating them from plant metabolites.

3.2. *Fungal Metabolites*

In this case, four *k*-NN QSAR models have been trained to separate 825 fungal metabolites (assigned 1.0 activity values) from the rest of the compounds that have been considered as inactive, with assigned 0.0 dependent variables.

3.3. *Bacterial Metabolites*

To study this system, 551 bacterial metabolites from the *AnalytiCon-Discovery* collection have been assigned 1.0 activity value and remaining 6957 general drugs, drug-likes, antimicrobials, fungal, plant and human metabolites all have been considered as a negative control and assigned to null dependent variable.

3.4. *Human Metabolites*

The dataset of 1102 chemical substances involved with chemical reactions taking place in human body have recently been catalogued by the group of Prof. Wishart at the University of Alberta. These molecules have been incorporated to the larger metabolomics database and have been made available through the web: http://www.metabolomics.ca/. Thus, we attempted developing 'Human-Metabolite-Likeness' QSAR model hoping that the corresponding QSAR classifiers may become useful tools for assessing potential human therapeutics. We trained the *k*-NN approach to recognize 1104 human metabolites among 7508 compounds under study.

3.5. *QSAR Modeling*

All four classification systems 3.1–3.4 have been investigated using 10 fold cross-validation approach. In particular, within four classification systems for Bacterial-, Fungal-, Plant- and Human metabolites, all 7508 substances have been separated into active and inactive components, according to the protocols described above, and then have been separated into ten 90%-10% training/testing sets (where the training sets do not overlap), and keeping the ratio of active and inactive entries constant.

At the next step 59 normalized QSAR descriptors have been used as independent variables to train *k*-NN based models.

Four classification systems 3.1-3.4 have been independently processed within the *k*-NN training procedure, as described in the previous section and the performance of the resulting QSAR models has been assessed by the combined

True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) predictions on the testing sets. The corresponding parameters have then been transformed into Sensitivity, Specificity and Accuracy values that can be found in Table 1.

Table 1. Cross-validation confusion matrices and accuracy parameters for the developed binary QSAR classifiers for Bacterial-, human-, Fungal-, and Plant metabolites.

Model	TP	FP	TN	FN	SEN	SPE	ACC
Bacterial Metabolites	298	303	6654	253	0.541	0.956	0.926
Human Metabolites	856	291	6115	246	0.777	0.955	0.928
Fungal Metabolites	498	322	6361	327	0.604	0.952	0.914
Plant Metabolites	2179	211	4946	172	0.927	0.959	0.949

The data in Table 1 illustrates, that the method of *k*-Nearest Neighbors allowed generally accurate separation of actives and inactives in all four systems 3.1-3.4. Thus, the use of 59 'inductive' and conventional QSAR descriptors allowed almost 95% accurate recognition of plant metabolites, followed by 92.8%, 92.6% and 91.4% accuracy estimated for Human-, Bacterial- and Fungal- metabolites respectively.

These results confirm good predictive power of 'inductive' QSAR descriptors that has been previously attributed to the fact that they cover a broad range of proprieties of bound atoms and molecules related to their size, polarizability, electronegativity, electronic and steric interactions and, thus, can adequately capture structural determinants of intra- and inter-molecular interactions [1-5].

3.6. *Cross recognition and Similarity between Metabolites, Antibiotics, Drugs and Drug-like Substances.*

Notably, with the exception of the model for classification of plant metabolites, all other QSAR approaches produced non-dismissible number of false positive predictions (see Table 1) determined by overlaps between the studied groups of compounds.

To further investigate the extend of cross-recognition between four groups of native metabolites we re-trained 3-NN models 3.1-3.4 leaving one of the activity groups out of consideration and then applied the developed models to the excluded set. The resulting numbers of positive predictions have been collected into Table 2 and transformed into the corresponding fractions of antimicrobials, Drugs, Drug-likes, Bacterial–, Plant-, Fungal- and Human-metabolites that have been recognized by the 'non-self' QSAR models.

Table 2. Fractions of the studied groups of compounds recognized as false positive predictions by the developed four QSAR models

	Bacterial Metabolites	**Human Metabolites**	**Fungi Metabolites**	**Plant Metabolites**
Antibacterials	3.5	1.7	3.1	2.9
Drugs	2.0	2.2	0.4	1.4
Chemicals	0.5	0.6	0.8	0.4
Bacterial Metabolites		0.7	31.4	0.9
Human Metabolites	1.4		7.4	8.0
Fungi Metabolites	22.1	3.8		11.3
Plant Metabolites	0.3	1.4	6.0	

These numbers reflect a profound similarity between Fungal and Bacterial metabolites as well as between Fungal and Plant metabolites (interestingly, no significant overlaps have been established for Plant and Bacterial substances). Human metabolites demonstrated no significant cross-recognition with other natural compounds which confirms the previously reported stand-alone nature of this class of substances.

When the developed QSAR 'metabolite-likeness' models have been applied to the groups of conventional human therapeutics, antibacterials and inactive drug-like chemicals, some interesting overlaps have been found between antibacterials and bacterial metabolites as well as between drugs and human metabolites (see the upper part of Table 2). More detailed analysis of substances recognized by the 'human metabolite-likeness' classifier demonstrate, that the largest portion of the corresponding false positive predictions originated by the fungal metabolite substances (likely reflecting strongest resemblance between fungal and human cellular composition), followed by natural molecules of plant origin and bacterial metabolites (as illustrated by Figure 2).

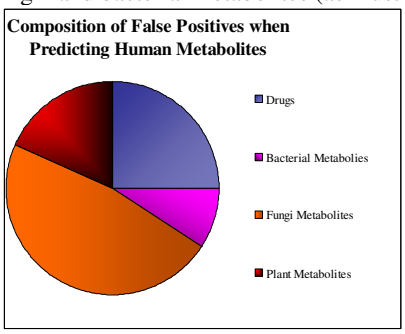

Figure 2. Composition of false positives produced by the QSAR model for Human metabolites.

Nonetheless, general overlap of Human metabolites with other studied groups of molecules is very limited. To illustrate positioning of Human metabolites against other groups in the chemical space we projected the

corresponding entries onto three Principal Components derived from 59 used QSAR descriptors (Figure 3)

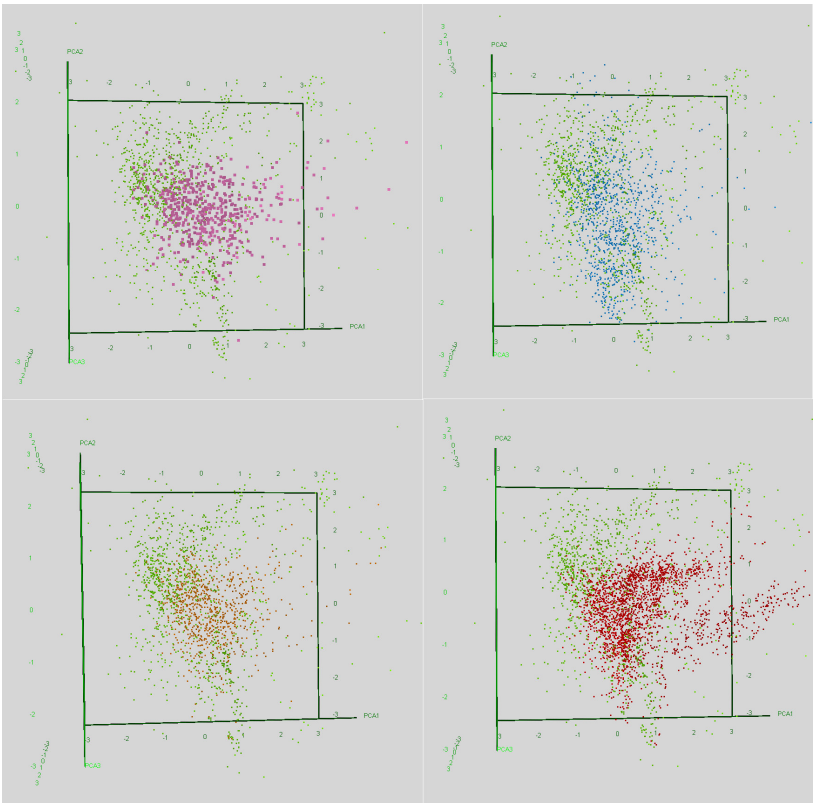

Figure 3. Separation of Human metabolites (Green) from other groups of the studied compounds in three dimensional space formed by 3 Principal Components derived from 59 used QSAR descriptors. The color coding of points corresponds to the following scheme: Red: Plant Metabolites, Orange: Fungal Metabolites, Pink: Bacterial Metabolites, Blue: Drugs.

The chart demonstrates that descriptors computed for human metabolites determine certain overlap between Human metabolites and other substances, but their overall positioning in the chemical space is quite distinguished and rather sparse compared to other cluster that are much more compact. This, likely, can be attributed to the most diverse nature of substances involved in the human chemical pathways.

4. Conclusions

To summarize the results of the previous sections, it is possible to conclude that antimicrobials, conventional therapeutics, inactive chemicals, as well as plant, fungal and bacterial metabolites are organized into rather compact as distinguished clusters in QSAR descriptors space what makes it possible to distinguish these types of chemicals with binary SA models. Fungal metabolites demonstrate rather significant mutual overlap with Bacterial substances and some degree of resemblance with Plant derivatives. When we utilized the *k*-Nearest Neighbors, algorithm for the purpose of recognizing four groups of metabolic substances it allowed their generally acceptable separation.

When the developed four 'metabolite-likeness' models have been applied to conventional human therapeutics and specific antimicrobial substances the formers demonstrated strongest association with human metabolites, while the later demonstrated tendency of 'bacterial metabolite – like' behavior. It is possible to speculate that the established 'drugs-human metabolites' and 'antimicrobials–bacterial metabolites' associations result from strict bioavailability requirements imposed on therapeutics which, in a way, enforce their metabolite-like properties.

The overall results of the conducted comparative QSAR analysis bring more insight into the nature and structural dominants of the studied classes of chemicals substances and, if necessary, can help rationalizing the design and discovery of novel antimicrobials and human therapeutics with metabolite-like chemical profiles.

Acknowledgments

Authors thank Dr. David Wishart (University of Alberta) for providing us the Database of Human Metabolites

Appendix

'Inductive' (from 1 to 26) and conventional QSAR parameters (from 27 to 59) used for creating binary QSAR models 3.1-3.4.

Average_EO_Neg, Average_EO_Pos, Average_Hardness, Average_Neg_Charge, Average_Neg_Hardness, Average_Pos_Charge, Average_Softness, EO_Equalized, Global_Softness, Hardness_of_Most_Neg, Hardness_of_Most_Pos, Largest_Neg_Hardness, Largest_Neg_Softness, Largest_Pos_Hardness, Largest_Rs_i_mol, Most_Neg_Rs_mol_i, Most_Neg_Sigma_i_mol, Most_Neg_Sigma_mol_i, Most_Pos_Charge, Most_Pos_Rs_i_mol, Most_Pos_Sigma_i_mol, Most_Pos_Sigma_mol_i,

Softness_of_Most_Pos, Sum_Hardness, Sum_Neg_Hardness, Total_Neg_Softness,

b_double, b_rotN, b_rotR, b_triple, chiral, rings, a_nN, a_nO, a_nS, FCharge, lip_don, KierFlex, a_base, vsa_acc, vsa_acid, vsa_base, vsa_don, density, logP(o/w), a_ICM, chi1v_C, chiral_u, balabanJ, logS, ASA, ASA+, ASA-, ASA_H, ASA_P, CASA+, CASA-, DASA, DCASA

For more details on 'inductive' parameters see references [1-5], while the used conventional QSAR parameters can be accessed through the MOE program [16].

References

1. A. Cherkasov, *Curr. Comp.-Aided Drug Design.* **1**, 21 (2005).
2. A. Cherkasov, and B. Jankovic, *Molecules.* **9**, 1034 (2004).
3. A. Cherkasov, Z. Shi, M. Fallahi, and G.L. Hammond, *J. Med. Chem.* **48**, 3203 (2005).
4. A. Cherkasov, *J. Chem. Inf. Model.* **46,** 1214 (2006).
5. E. Karakoc, S. C. Sahinalp, and A. Cherkasov. *J. Chem. Inf. Model.* **46,** *in press* (2006).
6. *ChemIDPlus* database: http://chem.sis.nlm.nih.gov/chemidplus/ , May 2006
7. *Journal of Antibiotics* database*:* http://www.nih.go.jp/~jun/NADB/byname.html , May 2006
8. F. Tomas-Vert, F. Perez-Gimenez, M.T. Salabert-Salvador, F.J. Garcia-March, J. Jaen-Oltra, *J. Molec. Struct. (Theochem).* **504**, 249 (2000).
9. M.T.D. Cronin, A.O. Aptula, J.C. Dearden, J.C. Duffy, T.I. Netzeva, H. Patel, P.H. Rowe, T.W. Schultz A.P. Worth, K. Voutzoulidis, and G. Schuurmann, *J. Chem. Inf. Comp. Sci*. **42**, 869 (2002).
10. M. Murcia-Soler, F. Perez-Gimenez, F.J. Garcia-March, M.T. Salabert-Salvador, W. Diaz-Villanueva, M.J. Castro-Bleda and A. Villanueva-Pareja. *J Chem Inf Comput Sci.* **44**, 1031 (2004).
11. The Merck Index 13.4 CD-ROM Edition, *CambridgeSoft*, Cambridge, MA, 2004.
12. *Analyticon Discovery* Company: www.ac-discovery.com May 2006
13. *Assinex Gold Collection*, Assinex Ltd., Moscow, 2004.
14. *Human Metabolome Database*: http://redpoll.pharmacy.ualberta.ca/~aguo/www_hmdb_ca/HMDB/ , May 2006
15. T.A. Halgren, *J. Comp. Chem.* **17**, 490 (1996).
16. *Molecular Operational Environment*, **2005**, by Chemical Computing Group Inc., Montreal, Canada.
17. E. Karakoc, A. Cherkasov, and S. C. Sahinalp. *Bioinformatics*, in press

(2006).
18. CPLEX: High-performance software for mathematical programming http://www.ilog.com/products/cplex/, May 2006.
19. M. Tasan, J. Macker, M. Ozsoyoglu, S. Cenk Sahinalp. Distance Based Indexing for Sequence Proximity Search, IEEE Data Engineering Conference ICDE'03, Banglore, India (2003)