ELSEVIER

Combinatorial networks

Victor S. Lobanov and Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc., Exton, PA, USA

A novel approach for the analysis and virtual screening of large combinatorial libraries is presented. The method attempts to relieve the computational burden by computing the properties of the products in a way that does not require their explicit enumeration. In particular, a small subset of compounds from the virtual library is identified and their descriptors are calculated in a conventional manner. The resulting data is used as input to a multilayer perceptron, which is trained to predict the descriptors of the products from the descriptors of their respective building blocks. Once trained, the neural network is able to estimate the descriptors of the remaining members of the virtual library with remarkable accuracy, without ever, generating their connection tables. This method eliminates the two most time-consuming steps in virtual screening and allows the processing of very large combinatorial libraries that are intractable with conventional techniques. © 2001 by Elsevier Science Inc.

Keywords: Combinatorial library, combinatorial chemistry, high-throughput screening, virtual screening, compound selection, molecular similarity, molecular diversity, molecular descriptor, descriptor prediction, neural network

INTRODUCTION

Algorithmic efficiency has been a long-standing objective in computational drug design. There is perhaps no other problem in chemistry where the need for efficiency is as pressing as in combinatorial chemistry. Whether it is based on molecular diversity, molecular similarity, or structure–activity correlation, the design of a combinatorial experiment usually involves the enumeration of every product in the virtual library, and the computation of key molecular properties that are thought to be pertinent to the application at hand.[1–3] In the past few years, several effective methodologies have been developed, some specific to the problem of diversity profiling and similarity searching,[4–7] and others of much broader applicability.[8–12] While product-based design is straightforward for relatively small collections, many combinatorial libraries can reach staggering sizes, which render them inaccessible by conventional techniques. In such cases, the most common solution is to restrict attention to a smaller subset of the virtual library, or to consider each substitution site independently of all the others.[13–14] This approach is intuitive to a chemist, and the resulting designs, being full combinatorial arrays, require minimal synthetic resources to execute and can be easily augmented if there are problems with synthesis or reagent availability. Unfortunately, it has been shown that this increase in performance is often done at the expense of the primary objective for which the experiment was designed.[16,17]

Recently, we proposed an efficient product-based similarity searching algorithm that does not require exhaustive enumeration of every product in the virtual library.[18] This approach is based on the notion that the structural diversity of a combinatorial library stems from a limited number of building blocks, and that it is possible, through random sampling, to identify the reagents that lead to the products that are most closely related to the query structure. As with any application of this kind, the results are critically dependent upon the choice of descriptors used to define chemical space and the definition of chemical distance. For all but the simplest cases, the calculation of molecular descriptors is a computationally intensive process that requires, at a minimum, knowledge of the product's molecular graph. We present an alternative approach that circumvents this requirement and can dramatically reduce the computational effort required for the analysis of large collections. This is accomplished using feed-forward neural networks, which are trained to predict the descriptors of products from the descriptors of their respective reagents (Figure 1). Once the networks are trained, screening the virtual library (or any subset thereof) becomes a matter of retrieving the precomputed descriptors of the reagents, feeding them through the neural networks to compute the descriptors of the products, and using these descriptors for any subsequent analysis, searching, or classification task. Unlike previous attempts, which focused exclusively on decomposable descriptors,[19,20] this method is general and can be applied to a wide variety of molecular properties, regardless of origin and complexity. The algorithm is fast, and permits the screening of virtual libraries at a rate of a few thousand compounds per second on a modern personal computer.

This article describes the general architecture of combinatorial neural networks (CNNs) and demonstrates their utility for similarity searching of large combinatorial libraries. The

Color Plates for this article are on pages 610–613.

Corresponding author: V.S. Lobanov, 3-Dimensional Pharmaceuticals, 665 Stockton Drive, Exton, PA19341, USA. Tel.: 610-458-5264, Ext. 6501; Fax: 610-458-8249.

E-mail address: victor@3dp.com (V.S. Labanov).

Journal of Molecular Graphics and Modelling 19, 571–578, 2001

655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(01)00093-6

method is tested on two well-known data sets and compared against the conventional approach involving exhaustive enumeration and evaluation of every product in the virtual library. The application of this technique to other important problems in combinatorial series design, including diversity profiling and structure–activity correlation, and its extension to other classes of descriptors will be presented elsewhere.[21]

METHODS AND DATA SETS

Virtual Libraries

The combinatorial libraries used in this study were taken from our previous work.[18] The first is a 4-component library based on the Ugi reaction: a set of 100 acids, 100 amines, 37 aldehydes, and 17 isonitriles were chosen at random from the Available Chemicals Directory (ACD),[22] and were used to generate a virtual library containing 6.29 million compounds. The second is a 3-component library based on the reductive amination reaction involving a diamine core and two sets of alkylating/acylating agents. A virtual library of 6.75 million compounds was generated using a set of 300 diamines and two sets of 150 alkylating/acylating agents selected at random from the ACD. The size of the two collections was intentionally restricted, so that an exhaustive search would be possible to validate the results obtained with our methodology.

Virtual libraries were generated using the enumeration classes of the Directed Diversity® toolkit.[23] These classes take as input lists of reagents supplied in SD or SMILES format, and a reaction scheme written in a proprietary Tcl-based scripting language[24] that uses SMARTS for substructure specification. All chemically feasible transformations are supported, including multiple reactive functionalities, different stoichiometries, cleavage of protecting groups, stereo-specificity, and many others. The computational and storage requirements of the algorithm are minimal (even a billion-membered library can be generated in a few seconds on a personal computer) and scale linearly with the number of reagents.

Molecular Descriptors

Both reagents and products were characterized by a well-established set of 117 topological descriptors, including molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev-Trinajstić indices, and topological state indices.[25,26] These descriptors have a long and proven track record in structure–activity analysis[25] can be computed directly from the connection table, and are consistent with the medicinal chemists' perception of molecular similarity. Moreover, they have been shown to exhibit proper 'neighborhood behavior'[27] and are thus well suited for diversity analysis and similarity searching.[18,28] These data were subsequently normalized and decorrelated using principal component analysis (PCA), resulting in an orthogonal set of 25 to 29 latent variables, which accounted for 99% of the total variance in the data. The PCA preprocessing step was necessary to eliminate duplication and redundancy in the data, which is typical of graph-theoretic descriptors.

For visualization purposes, this multidimensional data was further reduced to 2 dimensions using a very fast nonlinear mapping algorithm developed by our group.[29,30,31] The projection was carried out in such a way that the pair-wise distances between points in the multidimensional principal component space were preserved as much as possible on the 2-dimensional nonlinear map. The resulting projection was used to visualize the selections, which were carried out using all significant principal components.

Similarity Searching

Similarity searching represents the most common form of virtual screening. It is based on the 'similar property principle,'[32] i.e., the fundamental belief that structurally similar compounds tend to exhibit similar physicochemical and biological properties. Thus, given a set of compounds with some desired biological effect, one seeks to identify similar compounds, expecting that some of them will be more potent, more selective, or more suitable in some other way than the original leads. In the present study, the similarity between two compounds was measured by their Euclidean distance in the multidimensional space[33] formed by the principal components that preserved 99% of the variance in the original topological features.

Combinatorial Networks

Combinatorial networks are multilayer perceptrons (MLPs) trained to reproduce the descriptors of products from the descriptors of their respective building blocks (Color Plate 1). In the present implementation, each neural network comprised three layers: an input layer containing $r \times n$ neurons, where n is the number of reagent features and r is the number of variation sites in the combinatorial library; a hidden layer containing from 2 to 15 units depending on the complexity of the transformation; and an output layer having a single neuron for each product feature predicted by the neural network (Color Plate 1). Thus, each training sample consisted of two sets of descriptors: one or more features for each of the building blocks, and one or more features for the respective product. The reagent descriptors were concatenated into a single array, and were presented to the network in the same order (p_{11}, p_{12},. . ., p_{1n}, p_{21}, p_{22},. . ., p_{2n},. . ., p_{r1}, p_{r2},. . ., p_{rn}, where p_{ij} is the j-th descriptor of the reagent at the i-th variation site).

Our analysis was based on fully connected MLPs trained with the standard error back-propagation algorithm.[34] The logistic transfer function $f(x) = 1/(1 + e^{-x})$ was used for both hidden and output layers. Each network was trained for a fixed number of epochs or until a predefined error threshold was met, using a linearly decreasing learning rate from 1.0 to 0.01 and a momentum of 0.8. During each epoch, the training patterns were presented to the network in a randomized order. Initially, the training was constantly monitored with a separate validation set, but no signs of over-fitting were observed for any of the training sets and network parameters used in our study. Thus, this validation methodology was abandoned, and the performance of each neural network was evaluated after the system was trained, using a separate test set. A similar resistance to over-fitting was also observed in a related class of neural networks used for nonlinear mapping.[29]

Software

All computations, including virtual library generation, descriptor calculation, similarity searching, and network training, were carried out using proprietary software written in the C++

programming language and based on 3-Dimensional Pharmaceuticals' Mt++ class library.[23] These programs are part of the DirectedDiversity® software suite,[35] and were designed to run on all POSIX-compliant Unix and Windows platforms. Parallel execution on systems with multiple CPUs is supported through the multi-threading classes of Mt++. The calculations were performed on a Dell Dimension, workstation equipped with two 800 MHz Intel Pentium III processors running Windows 2000 Professional. Although most of the computations were carried out in multi-threaded mode, the reported times where scaled to a single processor.

RESULTS AND DISCUSSION

Network Performance

Three different architectures were examined in the present study: (1) networks that take as input a single descriptor from each reagent and produce a single descriptor for the product; (2) networks that take as input multiple descriptors from each reagent and produce a single descriptor for the product; and (3) networks that take as input multiple principal components from each reagent and produce a single principle component for the product. We will refer to the first category as single-input single-output (SISO) perceptrons, and the last two categories as multiple-input single-output (MISO) perceptrons. The performance of each architecture was evaluated using three statistical measures: the correlation coefficient between the actual and predicted descriptors, the amount of distortion of the similarity matrix as measured by Pearson correlation coefficient, and the effect of that distortion on similarity searching and context-based retrieval. Preliminary studies indicated that networks with multiple output nodes (i.e., multiple-input multiple-output [MIMO] perceptrons producing multiple product descriptors or principal components) were difficult to train and produced results that were inferior to those obtained with an ensemble of single-output networks. In the following discussion, we will use the term 'exact' to refer to descriptors computed with the conventional method, and the term 'approximate' to refer to the descriptors generated by the neural networks. The reader is reminded that this paper's main concern is algorithmic efficiency; the usefulness of molecular connectivity indices in diversity profiling, similarity searching, and QSAR/QSPR has been amply demonstrated in the past,[25] and will not be addressed in the present work.

The simplest of all architectures involves a series of networks, each of which is trained to predict the value of a single product descriptor from the values of the same descriptor of the corresponding reagents. Thus, for a library with r components, each descriptor is estimated by a SISO network with r input and 1 output nodes, hereafter denoted $r-h-1$, where h is the number of hidden nodes. This approach offers simplicity and ease of training, as well as access to the individual product descriptors. Unfortunately, we found that this method works well for only about 80% of the 117 topological descriptors used in our analysis (Table 1). The remaining 20% of the descriptors cannot be predicted reliably from the corresponding reagent descriptors alone. As with the other architectures, the predictive ability of the networks was assessed using an independent test set comprising 10,000 randomly chosen products that were not 'seen' by the networks during training.

This situation can be improved by increasing the number of synaptic parameters and by adding to the training data other reagent descriptors that can provide additional information needed for successful prediction. This leads to a network topology of the form $r \times n-h-1$, where n is the number of input descriptors per reagent. The additional descriptors can be chosen in a variety of ways. For example, one could employ a feature selection algorithm similar to that used in step-wise regression analysis: try all possible pairs of descriptors and select the best pair, then try all possible triplets keeping the first two descriptors fixed and select the best triplet, and continue in the same manner until a predefined number of descriptors of error threshold is met. Fortunately, it turns out that this rather intensive algorithm is unnecessary, and excellent results can be obtained with the following heuristic approach.

First, the correlation coefficients between each reagent and each product descriptor are calculated, and a series of SISO networks are trained in the manner described in the previous paragraph. Then, for each product descriptor that cannot be adequately modelled (i.e., having a training R^2 less than 0.9), the two reagent descriptors that are most highly correlated to that product descriptor are added to the training data, and a new MISO network is trained. When applied to the Ugi library, this approach resulted in an array of neural networks that were able to predict all 117 descriptors with high accuracy for both the training and test sets (Table 1). The correlation coefficients between the actual and predicted descriptors ranged from 0.77 to 1.0, with the smaller values typically associated with the more complex properties such as the Bonchev-Trinajstić[26] information index $\overline{I}_D^C$ and the Kappa shaped index ${}^3\kappa\alpha$.[25]

The need to employ additional reagent information to successfully predict some of the product descriptors becomes more evident if one considers the nature of these properties and how they depend upon the respective properties of the building blocks. Consider, for example, the count of different atom types in a molecule (descriptor number 3 in Table 1). Two halogen substituted aldehydes and one halogen substituted diamine, all supplying three different types of atoms each (excluding hydrogen), can give a product where the count of atom types ranges from three to five depending on whether the halogens are all the same or all different. It is impossible to reliably predict such a descriptor without supplying additional information by means of other descriptors, and this is precisely the reason for the improved performance of MISO-type CNNs over SISO networks.

To assess the impact on molecular similarity, the optimized networks were used in a feed-forward manner to estimate the descriptors of all 6.29 million compounds in the Ugi library. These descriptors were subsequently decorrelated using the rotation matrix derived from the training set, and the Pearson correlation coefficient of the resulting pairwise Euclidean distances was computed. This statistic, which measures the correlation between the similarity coefficients computed with the two sets of descriptors (exact versus approximate), had a value of 0.99, indicating a nearly perfect reproduction. This accuracy was also reflected in the context of similarity searching, using 10 randomly chosen compounds from that library as leads (queries). In particular, the 1,000 most similar compounds to each of these leads were identified using the PCs derived from both the exact and approximate descriptors, and their similarity scores were compared and summarized in Color Plate 2. Note that to permit a direct comparison, the hit lists obtained with the approximate descriptors were fully enumerated, and their

Table 1. Prediction of individual descriptors by single-output neural networks. MISO values are reported only for those descriptors that could not be adequately modeled by SISO networks

Index	Descriptor	SISO Training R^2	SISO Test R^2	MISO Training R^2	MISO Test R^2
1	no. atoms	0.996	0.997		
2	no. bonds	0.995	0.996		
3	no. elements	0.603	0.614	0.822	0.823
4	molecular weight	0.996	0.997		
5	chi 0	0.996	0.997		
6	chi path 1	0.996	0.997		
7	chi path 2	0.994	0.995		
8	chi path 3	0.971	0.973		
9	chi path 4	0.974	0.976		
10	chi path 5	0.956	0.957		
11	chi path 6	0.909	0.910		
12	chi path 7	0.837	0.843	0.943	0.942
13	chi path 8	0.666	0.673	0.938	0.934
14	chi path 9	0.563	0.554	0.939	0.936
15	chi path 10	0.447	0.457	0.950	0.950
16	chi cluster 3	0.988	0.987		
17	chi cluster 4	0.993	0.993		
18	chi path/cluster 4	0.978	0.980		
19	val chi 0	0.996	0.997		
20	val chi path 1	0.997	0.998		
21	val chi path 2	0.996	0.996		
22	val chi path 3	0.993	0.994		
23	val chi path 4	0.981	0.982		
24	val chi path 5	0.952	0.951		
25	val chi path 6	0.907	0.905		
26	val chi path 7	0.773	0.775	0.901	0.905
27	val chi path 8	0.619	0.621	0.890	0.889
28	val chi path 9	0.349	0.328	0.910	0.910
29	val chi path 10	0.222	0.201	0.921	0.920
30	val chi cluster 3	0.994	0.994		
31	val chi cluster 4	0.993	0.993		
32	val chi path/cluster 4	0.988	0.989		
33	chi chain 3	1.000	1.000		
34	chi chain 4	1.000	1.000		
35	chi chain 5	0.979	0.978		
36	chi chain 6	0.995	0.995		
37	chi chain 7	0.999	0.999		
38	chi chain 8	1.000	1.000		
39	chi chain 9	0.999	0.999		
40	chi chain 10	0.999	0.998		
41	val chi chain 3	1.000	1.000		
42	val chi chain 4	1.000	1.000		
43	val chi chain 5	0.994	0.996		
44	val chi chain 6	0.994	0.995		
45	val chi chain 7	0.998	0.998		
46	val chi chain 8	1.000	1.000		
47	val chi chain 9	0.997	0.998		
48	val chi chain 10	0.986	0.980		
49	subgraph count path 2	0.996	0.997		
50	subgraph count path 3	0.990	0.990		
51	subgraph count path 4	0.957	0.960		
52	subgraph count path 5	0.914	0.918		

(Continued)

Table 1. ***(Continued)***

Index	Descriptor	SISO Training R^2	SISO Test R^2	MISO Training R^2	MISO Test R^2
51	subgraph count path 4	0.957	0.960		
52	subgraph count path 5	0.914	0.918		
53	subgraph count path 6	0.837	0.844	0.909	0.905
54	subgraph count path 7	0.752	0.770	0.892	0.887
55	subgraph count path 8	0.582	0.599	0.907	0.906
56	subgraph count path 9	0.446	0.448	0.933	0.932
57	subgraph count path 10	0.366	0.383	0.947	0.945
58	subgraph count cluster 3	0.994	0.995		
59	subgraph count cluster 4	0.991	0.991		
60	subgraph count path/cluster 4	0.980	0.980		
61	subgraph count ring 3	1.000	1.000		
62	subgraph count ring 4	1.000	1.000		
63	subgraph count ring 5	0.995	0.995		
64	subgraph count ring 6	0.994	0.995		
65	subgraph count ring 7	1.000	1.000		
66	subgraph count ring 8	1.000	1.000		
67	subgraph count ring 9	1.000	1.000		
68	subgraph count ring 10	0.999	0.999		
69	kappa 0	0.980	0.980		
70	kappa 1	0.991	0.992		
71	kappa 2	0.907	0.908		
72	kappa 3	0.709	0.710	0.806	0.799
73	kappa alpha 1	0.987	0.987		
74	kappa alpha 2	0.895	0.897	0.960	0.955
75	kappa alpha 3	0.685	0.686	0.774	0.770
76	Wiener path no.	0.967	0.965		
77	total Wiener path no.	0.903	0.892		
78	Shannon Index	0.911	0.911		
79	total no. of paths	0.939	0.932		
80	Bonchev-Trinajstić IdW	0.958	0.955		
81	Bonchev-Trinajstić mean IdW	0.972	0.972		
82	Bonchev-Trinajstić IdC	0.979	0.978		
83	Bonchev-Trinajstić mean IdC	0.793	0.773	0.777	0.759
84	Wiener parity no.	0.988	0.989		
85	Platt F no.	0.996	0.997		
86	delta partition 1	0.996	0.996		
87	delta partition 2	0.992	0.992		
88	delta partition 3	0.997	0.997		
89	delta partition 4	0.995	0.996		
90	delta partition 5^1	1.000	1.000		
91	delta partition 6^1	1.000	1.000		
92	no. H	0.996	0.997		
93	no. B^1	1.000	1.000		
94	no. C	0.997	0.998		
95	no. N	0.995	0.995		
96	no. O	0.994	0.993		
97	no. F	0.996	0.996		
98	no. Si[a]	1.000	1.000		
99	no. P	0.999	0.999		
100	no. S	0.997	0.999		
101	no. Cl	0.997	0.997		
102	no. Ge[a]	1.000	1.000		
103	no. As[a]	1.000	1.000		
104	no. Se[a]	1.000	1.000		

(Continued)

Table 1. ***(Continued)***

Index	Descriptor	SISO Training R^2	SISO Test R^2	MISO Training R^2	MISO Test R^2
107	no. halogens	0.997	0.998		
108	total topological state 1	0.924	0.918		
109	total topological state 2	0.947	0.945		
110	total topological state 3	0.904	0.888		
111	total topological state 4	0.956	0.956		
112	total topological state 5	0.852	0.826	0.915	0.907
113	total topological state 6	0.980	0.980		
114	total topological state 7	0.832	0.790	0.914	0.898
115	total topological state 8	0.988	0.988		
116	total topological state 9	0.913	0.909		
117	total topological state 10	0.922	0.918		

similarity scores were reevaluated using 'exact' descriptors computed in the conventional, direct manner. As shown in Color Plate 2, in all 10 cases, the two designs had nearly identical scores and very similar content, with an overlap ranging from 75 to 86% (Table 2). The equivalence of these selections for one of the leads is graphically illustrated in the nonlinear map in Color Plate 3. The entire screening process, including enumeration of the training set, network training, decorrelation, and similarity searching, required only 35 min of CPU time, which represents a 30-fold improvement in throughput compared with the direct approach.

Since principal components are often the desired output, significant improvements could be achieved if the evaluation of the individual descriptors were circumvented and the combinatorial networks were trained to predict the principal components directly. As we mentioned before, high-dimensional data sets are almost always redundant; in the case at hand, the 117 topological descriptors can be reduced to 25–30 latent variables without any significant loss in the contribution to variation. The presence of correlated variables affects molecular similarity in two important ways: redundant features are effectively taken into account with a higher weight, and there is a substantial and unnecessary increase in the computational effort required for data analysis.

The methodology involved the following steps. A sample set of 10,000 compounds was selected at random from the entire Ugi library, and was characterized using our standard set of 117 topological descriptors. These descriptors were normalized and decorrelated to 25 principal components, which accounted for 99% of the total variance in the data. In addition, all the reagents involved in making the entire Ugi library were described by the same set of descriptors, and were independently normalized and decorrelated to 27 principal components using the same variance cutoff. These data were used to develop an array of 25 CNNs (denoted PC-MISO), each of which was trained to predict one of the product PCs using all 27 PCs from each of the 4 input reagents. Thus, each neural network comprised 108 input, 2 hidden, and 1 output neurons (experiments showed that increasing the number of hidden neurons beyond 2 did not offer any significant improvements in the predictive ability of the resulting networks). A set of 10,000 input–output pairs was randomly split into a training set containing 90% of the samples and a test set containing the remaining 10% of the samples, and each neural network was trained on the training set for 100 epochs or until a predefined error threshold was met. Once training was complete, the combinatorial networks were used in a feed-forward manner to predict the 25 PCs for all 6.29 million compounds in the Ugi library, which were, in turn, used to identify the 1,000 most similar compounds to each

Table 2. Average similarity scores and percent identity of the 1,000 most similar compounds selected from the 6.29 million-member Ugi library. Identity refers to the percentage of compounds in common between the similarity-based selections produced using estimated and real descriptors

Lead	Random Similarity	Direct Similarity	SISO/MISO Similarity	SISO/MISO Identity	PC-MISO Similarity	PC-MISO Identity
1	1.754	0.480	0.501	69%	0.486	86%
2	1.158	0.238	0.279	56%	0.244	83%
3	1.664	0.655	0.680	64%	0.660	84%
4	1.291	0.179	0.213	60%	0.186	76%
5	1.763	0.327	0.335	82%	0.334	83%
6	1.196	0.201	0.224	58%	0.209	75%
7	1.294	0.274	0.291	72%	0.283	77%
8	1.385	0.268	0.288	73%	0.275	84%
9	1.694	0.464	0.481	74%	0.470	86%
10	1.613	0.460	0.470	79%	0.464	87%

of the 10 leads described above. The obtained selections were finally assessed using 'exact' PCs and compared with the ideal solutions (Color Plate 2). Again, in all 10 cases the selections were very similar to those derived with 'exact' descriptors and slightly better than those derived with regular SISO and MISO CNNs, both in terms of their similarity scores and the identity of the selected compounds, which ranged from 80 to 85% (Table 2). The entire screening process required only 39 min on the same 800-MHz Intel Pentium III processor.

To validate the generality of our approach, similar searches were carried out for the 3-component diamine library, using the same set of 117 topological descriptors for both reagents and products. In this case, 29 and 28 PCs were necessary to capture 99% of the variance in the reagent and products descriptors, respectively. Thus, 3-3-1 SISO and 9-3-1 MISO networks were used to predict individual descriptors, and 87-3-1 PC-MISO networks were employed for the prediction of principal components. As with the Ugi library, 10 leads were selected at random from the entire library and the 1,000 most similar compounds to each of these leads were identified using the PCs derived from both the exact and approximate descriptors. Once again, the selections obtained with approximate PCs were virtually identical to the ideal solutions, with PC-MISO predictions leading to slightly better similarity scores (Color Plate 4).

Of course, the most important aspect of this work is the impressive gain in performance afforded by the neural networks (Color Plate 5). In both cases, the similarity searches using exact descriptors required approximately 20 CPU hours, whereas the searches involving CNNs were completed in less than 40 minutes, which represents more than a 30-fold increase in performance. Although for both libraries the training of SISO, MISO, and PC-MISO CNNs required comparable execution times, the latter performed slightly but consistently better. On the other hand, SISO and MISO networks provide access to individual descriptors, which may have additional utility in applications such as diversity profiling. ADME modeling, and structure–activity correlation.

Composition of Training Set

A common concern with any machine learning algorithm is its dependence on the nature of the training set. To examine the effect of the composition of the training set on the quality of the predictions obtained by CNNs, 10 random samples of 10,000 compounds were drawn from the Ugi library and were used to train 10 different sets of 25 PC-MISO networks. The average R^2 between the pairwise distances computed with 'exact' and 'approximate' PCs over all 10 trials was 0.9951 ± 0.0004 and 0.9951 ± 0.0006 for the training and test set, respectively. The R^2 was computed by comparing the Euclidean distances between 1,000,000 randomly chosen pairs of compounds in the two PC spaces (exact versus approximate). Similar standard deviations were also observed for the diamine library (0.0003 and 0.0007 for the training and test set; see Color Plate 6), which suggests that the training of CNNs is both stable and convergent.

Size of Training Set

The size of the training set has a moderate effect on the quality of predictions as long as it remains large enough to sample each reagent sufficiently. The predictions improve as the size of the training set increases, and eventually plateaus after a few thousand samples (Color Plate 7). For the Ugi library there was virtually no improvement in prediction when the size of the training set was doubled from 10,000 to 20,000 compounds, but this was not the case for the diamine library where the difference in R^2 was small but still noticeable. The reason is almost certainly related to the difference in the number of reagents involved in the construction of these libraries (254 for the Ugi and 400 for the diamine library), and the fact that, for a given sample size, each individual reagent is more extensively sampled in the Ugi library. The only disadvantage of using larger training sets is that longer times are required for descriptor calculation and network training. In general, the sample size should be determined by weighing the benefits of higher accuracy against the increasing cost of computation.

CONCLUSION

A novel approach for the analysis and virtual screening of large combinatorial libraries has been presented. By circumventing enumeration and replacing descriptor evaluation with a simple feed-forward pass through a multilayer perception, the method permits the *in silico* characterization and screening of large combinatorial libraries that are intractable by direct techniques. Although the descriptors produced by this algorithm are not 'exact,' the error is barely noticeable and has minimal impact on similarity searching. The method is more than an order of magnitude faster than conventional enumerative similarity searching methodologies, and this differential increases with the size and combinatorial complexity of the virtual library.

ACKNOWLEDGMENTS

The authors are indebted to Dr. Raymond F. Salemme of 3-Dimensional Pharmaceuticals, Inc. for his insightful comments and support of this work.

REFERENCES

1 Agrafiotis, D.K. The diversity of chemical libraries. In; *The Encyclopedia of Computational Chemistry*, Schleyer, P.v.R., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer, H.F. III, and Schreiner, P.R., Eds., Wiley, Chichester, 1998, 742–761

2 Agrafiotis, D.K., Myslik, J.C., and Salemme, F.R. Advances in diversity profiling and combinatorial series design. *Mol. Diversity* 1999, **4**, 1–22

3 Agrafiotis, D.K., Lobanov, V.S., Rassokhin, D.N., and Izrailev, S. The measurement of molecular diversity. In: *Virtual Screening of Bioactive Molecules*, H.-J. Böhm, H.-J., and Schneider, G., Eds., Wiley-VCH, Weinheim, 2000, pp. 265–300

4 Lajiness, M.S. An evaluation of the performance of dissimilarity selection. In:*QSAR: Rational Aproaches to the Design of Bioactive Compounds*, Silipo, C., and Vittoria, A., Eds., Elsevier, Amsterdam, 1991, pp. 201–204

5 Pearlman, R.S., and Smith, K.M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 28–35

6 Stanton, R.V., Mount, J., and Miller, J.L. Combinatorial library design: maximizing model-fitting compounds

with matrix synthesis constraints. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 701–705
7 Agrafiotis, D.K., and Lobanov, V.S. Ultrafast algorithm for designing focused combinatorial arrays. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 1030–1038
8 Agrafiotis, D.K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 841–851
9 Agrafiotis, D.K., and Lobanov, V.S. An efficient implementation of distance-based diversity metrics based on k-d trees. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 51–58
10 Gillett, V.J., Willett, P., Bradshaw, J., and Green, D.V.S. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.* 1999, **39**, 169–177
11 Rassokhin, D.N., and Agrafiotis, D.K. Kolmogorov-Smirnov statistic and its applications in library design. *J. Mol. Graphics Modell.*, in press.
12 Agrafiotis, D.K. Multiobjective optimization of combinatorial libraries, *IBM Systems Journal*, in press
13 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K., and Moos, W.H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery, *J. Med. Chem.* 1995, **38**, 1431–1436
14 Martin, E.J., Spellmeyer, D.C., Critchlow, R.E. Jr., and Blaney, J.M. Does combinatorial chemistry obviate computer-aided drug design? In:*Reviews in Computational Chemistry*, Vol. 10, Lipkowitz, K.B., and Boyd, D.B., Eds., VCH, Weinheim, 1997, pp. 75–100
15 Martin, E., and Wong, A. Sensitivity analysis and other improvements to tailored combinatorial library design. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 215–220
16 Gillet, V.J., Willett, P., and Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries *J. Chem. Inf. Comput. Sci.* 1997, **37**, 731–740
17 Jamois, E.A., Hassan, M., and Waldman, M. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 63–70
18 Lobanov, V.S., and Agrafiotis, D.K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 460–470
19 Downs, G.M., and Barnard, J.M. Techniques for generating descriptive fingerprints in combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 59–61
20 Cramer, R.D., Patterson, D.E., Clark, R.D., Soltanshahi, F., and Lawless, M.S. Virtual compound libraries: a new approach to decision making in molecular discovery research. *J. Chem. Inf. Comput. Sci.* 1998, **38**, 1010–1023
21 Lobanov, V.S., and Agrafiotis, D.K. manuscript in preparation
22 Marketed by MDL Information Systems, Inc., 140 Catalina Street, San Leandro, CA 94577.
23 Copyright © 3-Dimensional Pharmaceuticals, Inc., 1994–2000
24 Ousterhout, J.K. *Tcl and the Tk toolkit*, Addison-Wesley, New York, 1994
25 Hall, L.H., and Kier, L.B. The molecular connectivity chi indexes and kappa shape indexes in structure-property relations. In:*Reviews of Computational Chemistry*, Boyd, D.B., and Lipkowitz, K.B., Eds., VCH Publishers, New York, 1991, pp. 367–422
26 Bonchev, D., and Trinajstic N. Information theory, distance matrix and molecular branching, *J. Chem. Phys.*1977, **67**, 4517–4533
27 Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Weinberger, L.E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors, *J. Med. Chem.* 1996, **39**, 3049–3059
28 Lewis, R.A., Mason, J.S., and McLay, I.M. Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *J. Chem. Inf. Comput. Sci.* 1997, **37**, 599–614
29 Agrafiotis, D.K., and Lobanov, V.S. Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* 2000, **40**, 1356–1362
30 Rassokhin, D.N., Lobanov, V.S., and Agrafiotis, D.K. Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comp. Chem.*, in press
31 Agrafiotis, D.K., Rassokhin, D.N., and Lobanov, V.S. Multidimensional scaling of large molecular similarity tables. *J. Comp. Chem.*, in press
32 Johnson, M.A., and Maggiora, G.M. *Concepts and applications of molecular similarity*, Wiley, New York, 1990
33 Willett, P., Barnard, J.M., and Downs, G.M. Chemical similarity searching, *J. Chem. Inf. Comput. Sci.* 1998, **38**, 983–996
34 Haykin, S. *Neural Networks*, Macmillan, New York, 1994
35 Agrafiotis, D.K., Bone, R.F., Salemme, F.R., and Soll, R.M. System and method for automatically generating chemical compounds with desired properties, United States Patents 5,463,564, 1995; 5,574,656, 1996; 5,684,711, 1997; and 5,901,069, 1999