ELSEVIER

Computational mapping tools for drug discovery

Yan A. Ivanenkov[1], Nikolay P. Savchuk[1], Sean Ekins[2,3,4] and Konstantin V. Balakin[1,5]

[1] ChemDiv Inc., 6605 Nancy Ridge Drive, San Diego, CA 92121, USA
[2] Collaborations in Chemistry, Jenkintown, PA 19046, USA
[3] Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, Piscataway, NJ 08854, USA
[4] Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD, USA
[5] Institute of Physiologically Active Compounds, Russian Academy of Sciences, 142432 Severnyi Proezd 1, Chernogolovka, Noginsk Area, Moscow Reg., Russia

During the past decade, computational technologies have become well integrated in the modern drug design process and have gained in influence. They have dramatically revolutionized the way in which we approach drug discovery, leading to the explosive growth in the amount of chemical and biological data that are typically multidimensional in structure. As a result, the irresistible rush towards using computational approaches has focused on dimensionality reduction and the convenient representation of high-dimensional data sets. This has, in turn, led to the development of advanced machine-learning algorithms. In this review we describe a variety of conceptually different mapping techniques that have attracted the attention of researchers because they allow analysis of complex multidimensional data in an intuitively comprehensible visual manner.

Introduction

Humans can visualize quite complex data to differing degrees depending upon individual memory. Probably since prehistoric times, however, humans have also relied on maps to visualize very complex coordinates and topologies, as well as their relationship to the world. For example, surviving early maps relate to the earth (*c*1300) and the universe (1092) in 2D, maps on the very largest scale. It is only relatively recently that we have turned our attention to mapping the universe at the molecular scale and specifically determining the different molecules that inhabit this 'space' [1–3]. Starting with the molecules themselves is just the beginning, as one would also need to consider the physicochemical properties and the interactions with different biological systems, an incredibly complex and overwhelming amount of information. As we shall see some advanced mapping methods derive from our understanding of the neural networks involved in image perception by the primary visual cortex of the human brain.

Corresponding author: Balakin, K.V. (balakin@ipac.ac.ru)
URL: http://www.ipac.ac.ru

It is therefore no surprise that modern technologies of drug design and development confront us with a wealth of experimental and theoretical data, such as expression profiles of different genes from the human genome, data obtained from high-throughput biological screening of combinatorial libraries and data from different chemical or biological databases (e.g. ChemBank, DrugBank, HMDB, PDSP, PubChem, ZINC, CDD, eMolecules, ChemSpider, and so on). Extracting the useful knowledge deeply embedded in the complex array of chemical and biological information is a key task of modern computer-based informatics. In most cases such data are generally represented by highly structured information content, which is usually hidden under the complex set of intrinsic relationships or high-dimensional abstractions. Consequently there is a need for advanced and generally quite sophisticated graphical tools aimed at comprehensive visualization and structure–activity analysis of such high-dimensional data sets. These methods play an important role in gaining an intuitive understanding of complex and often contradictory relationships within multidimensional space. Advanced computational approaches targeted at dimensionality reduction and self-organizing mapping represent powerful tools for modern drug design and development.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2009.05.016

This review focuses specifically on nonlinear dimensionality reduction techniques, which are commonly based on advanced mapping and 'quasi-mapping' algorithms. It describes both basic theoretical principles and some practical applications of these approaches in the field of chemical data mining and visualization. The main mapping techniques described herein include clustering, diffusion maps (DM), self-organizing Kohonen maps (Kohonen SOMs), nonlinear sammon maps, stochastic proximity embedding (SPE) and IsoMap. Because of space limitations, the review does not cover several mapping algorithms, such as kernel maps [4], conformal eigenmaps (CEM) [5] and FastMap [6], which can be considered as extensions of the techniques listed above.

Computational mapping techniques for drug design

A variety of advanced computational algorithms and methods have been effectively applied recently in medicinal chemistry for dimensionality reduction and visualization of the chemical data of different types and structure. The majority of these computational models are commonly based on the basic principles of dimensionality reduction and mapping. In turn, dimensionality reduction is an essential computational technique for the analysis of a large-scale, streaming and tangled data. Typically these data are tightly packed in the multidimensional set of various numerical descriptors that cannot be properly analyzed by simple statistical procedures. Thus, dimensionality reduction providing a low-dimensional data representation (in the case of 2D projection or 3D projection it produces maps) plays an important role in gaining an intuitive understanding of the complex and often contradictory relationships dispersed chaotically within multidimensional data sets.

The classical methods of dimensionality reduction, for example, principal component analysis (PCA) and multidimensional scaling (MDS) are not specifically adapted for large data sets and 'straight' mapping. Therefore, there is a growing interest in novel soft-computing approaches that might be applicable to the analysis of such data sets providing a comprehensive visualization. For this purpose, advanced computational mapping techniques, such as agglomerative hierarchical clustering, which is primarily based on 2D-structural similarity measurement, self-organizing mapping, such as Kohonen SOMs, Sammon mapping and SPE algorithm, as well as generative topographic mapping and truncated Newton optimization strategy could be effectively used. The practical uses of some mapping techniques are, however, weakened by a 'quadratic' restriction relating to large data sets, but they have several key advantages compared to PCA and MDS. In contrast to the traditional linear PCA and MDS, advanced data mining techniques compute the principal eigenvectors of the kernel-based matrix/function, rather than those of the covariance matrix. A curious property of nonlinear mapping is that it preserves the global or relative topology of the original input vector space as close as possible in a faithful and unbiased manner.

Clustering

Clustering is a common and simple computational technique broadly used to partition a set of estimated data points into groups (clusters), so that the objects in each group share common characteristics in accordance with the distance or similarity measure used [7]. This technique has already found numerous applications in different scientific areas including drug discovery. With respect to chemical data mining, it allows scientist to reduce significantly the complexity of large chemical data sets to a more manageable size. All clustering techniques can be roughly classified in two key categories: hierarchical, which partition the data by successively applying the same procedure to clusters formed during previous iterations or non-hierarchical, which determine clusters in a single step [8]. There are several well-known cluster techniques that have been successfully used in QSAR analysis, including the most popular: Ward's algorithm, a representative example of hierarchical methods, *k*-means algorithm is a good example of partition-based methods, and Jarvis–Patrick algorithm—representative example of nearest neighbor methods.

Many efforts have been made to visualize the results of hierarchical and non-hierarchical clustering based on graph drawing and tree layout algorithms. The most commonly used method, particularly targeted for diversity applications, is the Jarvis–Patrick algorithm. Several key algorithms for graph construction, such as tree visualization and navigation, including hierarchical clustering, have been comprehensively reviewed by Herman [9]. Strehl and Ghosh [10] also have evaluated many computational algorithms for visualizing non-hierarchical clusters including the similarity matrix plot. It should be noted that the classical dendrogram remains the most popular cluster visualization method. The basic limitation of radial space filling and linear tree visualizations, however, is the decreasing size and resolution of clusters with many nodes or when you drill deep down into the hierarchy. To overcome these challenges Agrafiotis *et al.* [11] have recently developed a new radial space-filling method for visualizing cluster hierarchies based on radial space-filling system and a nonlinear distortion function, which transforms the distance of a vertex from the focal point of the lens. This technique has been particularly applied for visualization and analysis of two different types of chemical data sets: chemically diverse combinatorial libraries (Figure 1a,b) [11] and the conformational space of small organic molecules (Figure 1c) [11,12].

Radial clustergrams represent an intuitively comprehensible method for graphical visualization of cluster hierarchies. In contrast to conventional dendrograms where the terminal nodes are usually arranged in a linear fashion, radial clustergrams organize the nodes circularly. A visual color-coding basis allows the scientist to clearly understand how the aggregate properties are propagated up the hierarchical structure, and a convenient zoom tool can be effectively used to magnify areas of particular interest without losing sight of the wider context (Figure 1a–c).

Among other modern clustering algorithms that should also be mentioned, the following methods have been applied in chemoinformatics for various applications: the Maximin Algorithm [12,13], Stepwise Elimination and Cluster Sampling [14], HookSpace [15], Minimum Spanning Trees [16], Graph Machines [17], Singular Value Decomposition (SVD) and Generalized SVD [18].

Nonlinear mapping techniques

Among the various dimensionality reduction techniques that have been recently described in the scientific literature, nonlinear and self-organizing neural-based mapping are unique alternative approaches, owing to their conceptual simplicity and unrivalled ability to reproduce effectively an intrinsic topology and the

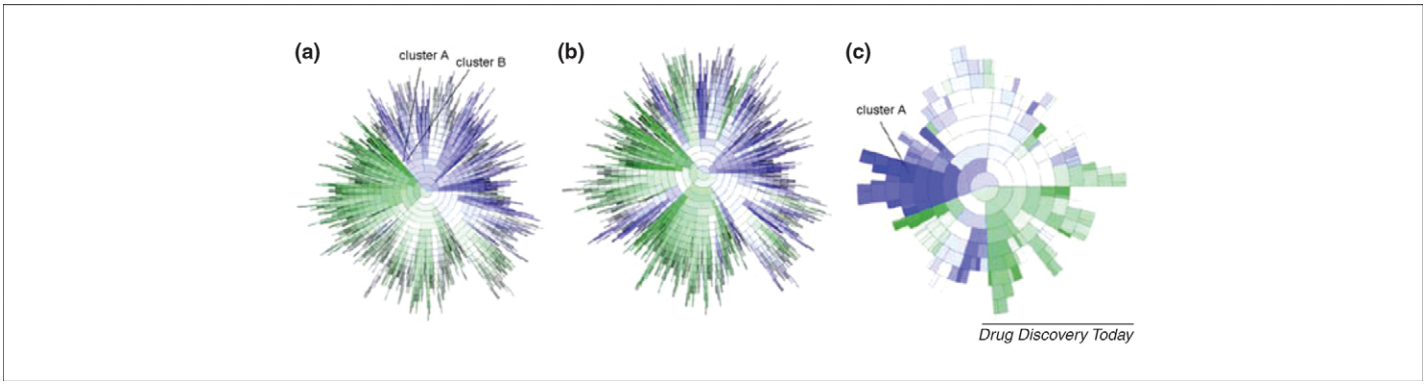

FIGURE 1

Radial clustergrams of a combinatorial library containing 2500 organic structures. The clustergrams, color-coded by the average molecular weight (a) and logP (b) (black color corresponds to high values of descriptors), highlighted by the two clusters designated as 'A' and 'B'. Although, these plots have a very similar appearance, which reflects a significant correlation between molecular weight and logP, it can be used to distinguish major families and subfamilies of molecules with related structures and properties at all levels of the hierarchy. Thus, sharp color changes across the cluster boundaries reveal structurally related chemical families with distinctly different properties. While all of the tested structures share a common topology, which explains their proximity in diversity space, compounds located in cluster 'A' contain several halogens as well as at least one bromine atom, which increases both their molecular weight and logP. There are no molecules in the first cluster with a bromine atom, and none of them carries more than one lighter halogen (F or Cl). (c) A radial clustergram color-coded by the radius of gyration (a measure of the extendedness or compactness of conformation) illustrates the application of the technique for visualizing of the conformational space around Amprenavir (HIV–protease inhibitor). Each pair of conformers was superimposed using a least-squares fitting procedure, and the resulting root-mean-square deviation (RMSD) was used as a measure of the similarity between two conformations.

hidden structure of the input data space in a faithful and unbiased manner.

Diffusion maps

Diffusion maps initially originated from the field of dynamical system theory. It is essentially a spectral clustering algorithm that is principally based on determination of the Markov random walk across the graph of input data [19]. Thus, the mapping from high-dimensional feature space to low-dimensional space (e.g. Euclidean space) evolves as the Markov chain progresses, where learning time controls the speed of diffusion. Following the algorithm, a specific measure of proximity between the input data points, also known as the diffusion distance, is implicitly defined through a number of time steps used. In a low-dimensional representation of the data studied, the pairwise diffusion distances are retained as well as possible to closely approximate the initial vector space. It should be especially noted that although diffusion maps are a global method, they can behave as a local method depending on the choice of the kernel function. The key underlying principle of diffusion distance lies in the large number of paths passing through the graph that makes the algorithm more robust to excessive noise level compared to, for example, the geodesic distance (see below). Although, diffusion maps are not widely used in drug design they have been successfully applied to gene expression analysis [20].

Self-organizing Kohonen maps

At least two different methods of self-organizing neural-based mapping are currently applied to dimensionality reduction, feature selection and topographic structure representation. Interestingly, the basic principle of the self-organizing methodology originates from studies related to the investigation of the mechanism of image perception translated by the primary visual cortex of the human brain. Willshaw and von der Malsburg were pioneers in this field who developed one of the first computational models in which artificial neurons were tightly packed into the two interrelated lattices (Figure 2a) [21].

As shown in Figure 2a, the 'input' lattice is projected onto the second two-dimensional plane by the corresponding synaptic route of the weight coefficients. The first lattice is simply constructed by the presynaptic neurons, while the second one consists of the postsynaptic neurons that are not formally assigned in accordance with the key principle—'winner takes all (WTA)'. Following both the short-range and long-range inhibitory mechanisms, neuron weights attached to the postsynaptic surface are adjusted iteratively following Hebb's learning rule until the optimal values are achieved. As a result, an increase in one synaptic weight directly leads to a decrease in others. Finally, it should be emphasized that the model is applicable solely for pattern recognition when the dimension of the input signal correlates closely to the dimension of the output feature image.

The second neural mapping methodology is based on the self-organizing strategy and is schematically outlined in Figure 2b. This approach was originally introduced by Kohonen in 1988 [22] and allows the construction of a low-dimensional topological representation of a high-dimensional data set by the optimal fixed amount of codebook feature vectors. In the Kohonen network, a learning process is firmly based on unsupervised logic and the target property is not considered within the training procedure. In contrast to supervised neural networks, SOM neurons are homogeneously arranged within the space spanned by a regular grid composed of many processing units in which the adaptation/learning process is generally performed by some predefined neighborhood rules. Primarily on the basis of Vector Quantization (VQ) strategy [23], these weight vectors are adjusted iteratively to the corresponding components of the input vector objects producing a visually understandable 2D or 3D-topological map. At the output, objects that are located close to each other in the input space should be closely embedded in the topologically isomorphic resulting space. Initially, all the Kohonen neurons receive

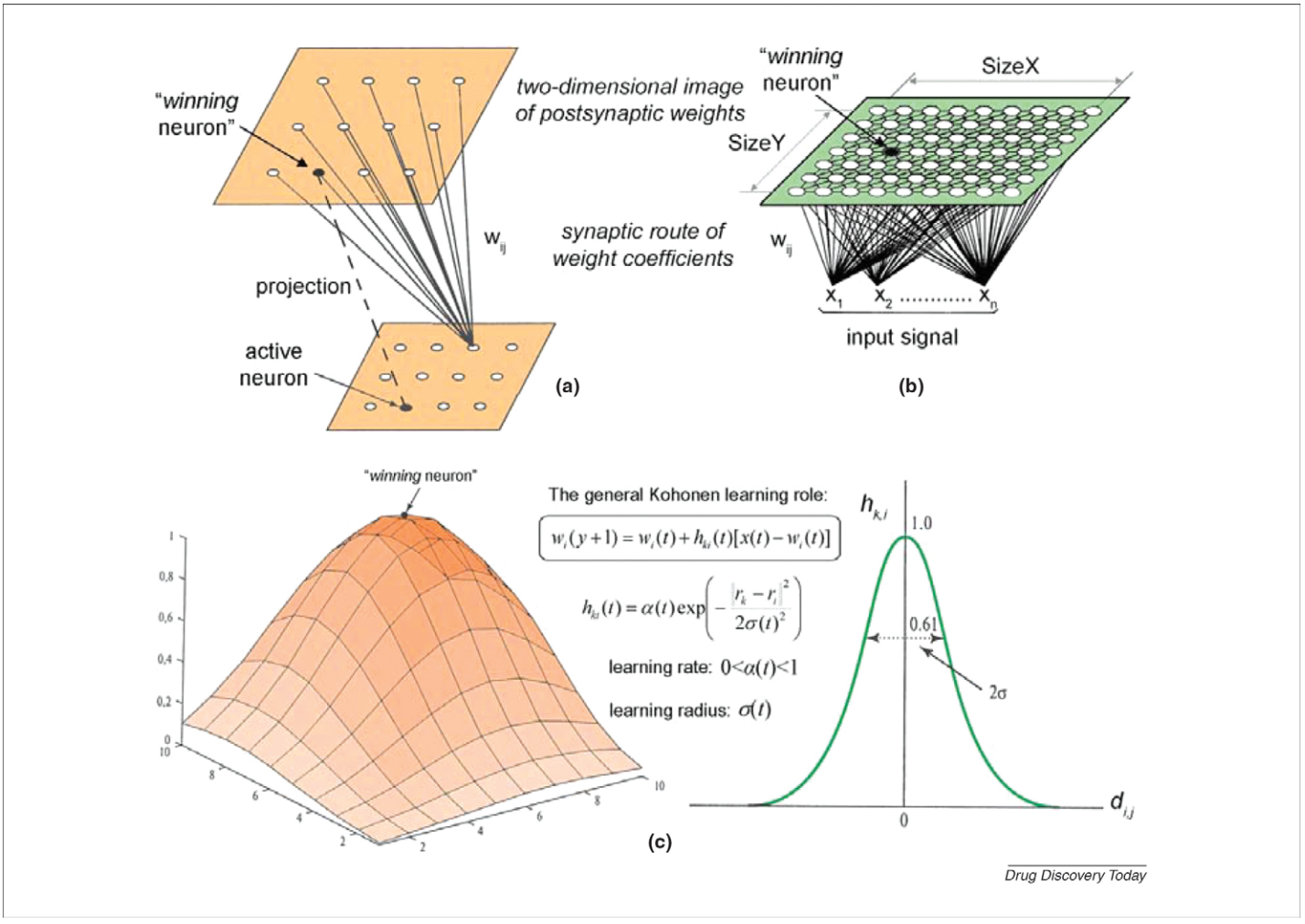

FIGURE 2

(a) The Willshaw–Malsburg's model of self-organizing mapping; (b) the Kohonen-based approach for dimensionality reduction and mapping; (c) a normal Gaussian distribution, one of the most widely applied neighborhood functions integrated in the Kohonen self-organizing algorithm.

identical input, and by means of lateral interactions they compete among themselves following the fundamental WTA principle. Typically, many training epochs are needed to complete the training process successfully. During the process, each neuron becomes peculiarly sensitive to a particular region of the original input space. The input samples that fall into the same region, whether they were or were not included in the original training set, are directly mapped onto the same neuron.

To date, a number of variants and different modifications of SOM have been developed and extensively applied in various scientific disciplines ranging from the engineering sciences to chemistry, medicine, biology, economics and finance. From a biological point of view, the Kohonen network is also biologically plausible in the same way as the Willshaw–Malsburg model [24]. From a functional point of view, the Kohonen algorithm accurately approximates the probability density functions calculated for each input variable by a finite set of reference vectors with the sole purpose of providing a low-dimensional data representation by using a nearest neighbor rule. To date, several nearest neighbor functions are commonly used to organize conceptually neurons within the Kohonen map faithfully to preserve initial multidimensional topology. These generally include the simplest 'Bubble neighborhood function' as well as multiparametric functions that describe the probability density distribution around the winning neuron. Following this approach, more than just one single winner neuron is adapted during each training cycle. Among these kernel-based functions the Gaussian (normal) distribution is routinely used in SOM providing a successful low-dimensional representation (Figure 2c).

Gasteiger *et al.* [25] were the first authors who successfully applied the SOM approach for the analysis and visualization of chemical data. Since then a huge number of scientific papers, dedicated to the application of self-organizing methodology in chemoinformatics, have been published. Several examples demonstrate how they can be used in drug discovery projects. Balakin *et al.* [26] have compiled a comprehensive reference database following a common experimental protocol on compound DMSO solubility (55 277 compounds with good DMSO solubility and 10 223 compounds with poor DMSO solubility), calculated specific physicochemical molecular descriptors (topological, electromagnetic, charge, and lipophilicity parameters), and, finally, applied two machine-learning approaches for training neural networks to address the DMSO solubility. Both the supervised (feed-forward, back-propagated neural networks) and unsupervised

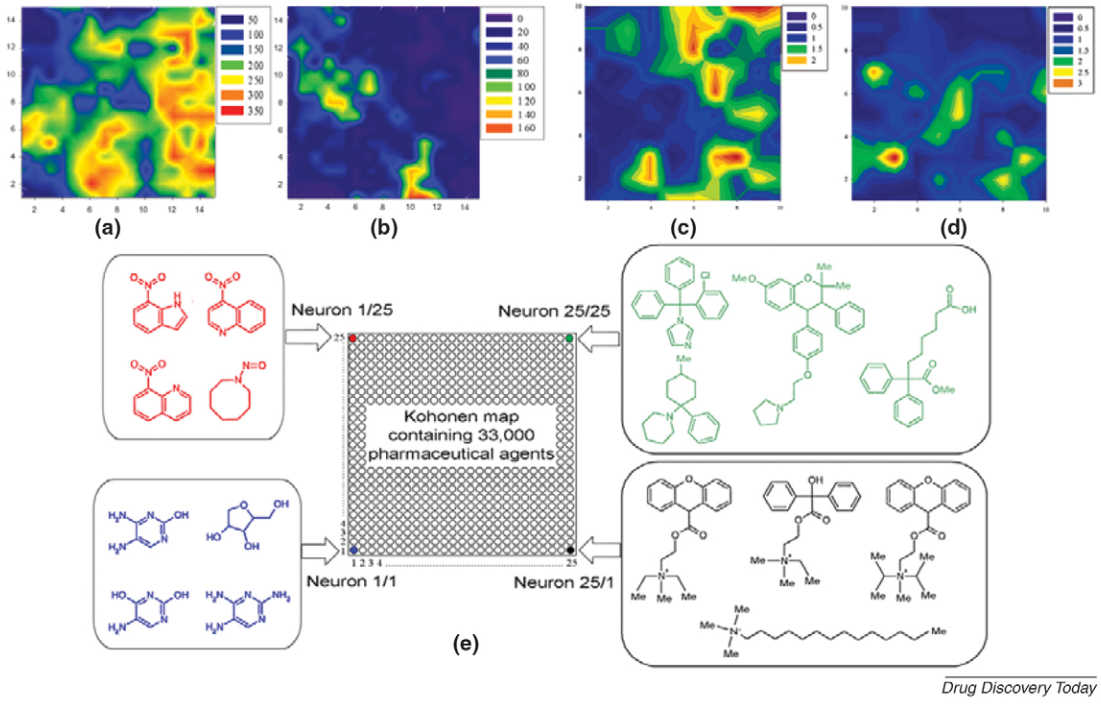

FIGURE 3

The distribution of DMSO(+) (a) and DMSO(−) (b) compounds within the Kohonen map; the Kohonen network developed for the prediction of cytochrome P-450 substrates (c) and products (d) (the distribution/density of compounds spread within the map is color-coded as follows: blue color is assigned to the low amount of compounds per node, while red color corresponds to the high value) (e) the Kohonen map (25 × 25) containing more than 33 000 structures form the MedChem database. The resulting map reflects the chemical feature space of the database used. The four representative groups of molecular structures are shown that are most similar to the weight vectors of the neurons (1/1, shown in blue), (1/25, shown in red), (25/1, shown in green), and (25/25, shown in black).

(Figure 3a,b) (self-organizing Kohonen neural networks) learning strategies were used. The resulting models were then externally validated by successfully predicting DMSO solubility of compounds in an independent test set. Korolev *et al.* [27] have also applied a self-organizing approach for successful computational modeling of cytochrome-mediated metabolic reactions (Figure 3c,d). A training database consisted of many known human cytochrome P-450 substrates (485 compounds), products (523 compounds), and non-substrates for 38 enzyme-specific groups (total of 2200 compounds) was compiled, and most typical cytochrome-mediated metabolic pathways within each group, as well as the substrates and products of these routes, were determined.

A further example demonstrates the usefulness of Kohonen-type SOMs for clustering massive chemical data sets (Figure 3e). Using the MedChem compound library as a source of pharmaceutically active substances (more than 33 000 entries were selected) a Kohonen network containing 25 × 25 output neurons arranged in a 2D-plane was trained and subsequently validated by Schneider and Wrede [28]. During the learning procedure, the network adapted the topology of the data distribution and, as a consequence, similar molecules were grouped together. This can be helpful to assess molecular diversity [29] and to construct small sets of compounds covering a defined variety of chemical features [30,31]. Following this concept, new endothelin antagonists were found using a Kohonen network for clustering known endothelin receptor ligands and subsequent database screening [32]. A similar approach was followed by Polanski [33] for identification of active site features in corticosteroid and testosterone binding globulins (CBG/TBG): self-organizing networks were developed for pseudo-receptor modeling based on a set of steroids with known CBG/TBG affinity data. Additional studies include those that focused on specific metabolic reactions, individual enzymes [34,35], receptors transporters and ion channels [36–39].

The practical relevance of Kohonen-based SOMs in the field of chemical data mining is quite obvious. According to the literature data, such computational models can be effectively applied at the earliest stages of drug discovery as a powerful tool for assessment/prediction of a number of key drug features including ADME and toxicity properties (e.g. blood–brain barrier (BBB) permeability, human intestinal absorption (HIA), half-life time ($t_{1/2}$) and volume of distribution (V_d) of drug compounds in human plasma, etc.), target-focusing specification (e.g. the common drug likeness *in silico* filter as well as GPCR-targeted, kinase-targeted, CNS-targeted libraries design, and so on), metabolic stability and pathways (for example, cytochrome P450 metabolism assessment). In addition, the algorithm was shown to be very effective during lead optimization and a broad chemical space analysis. For example, the modern conception of diversity-oriented synthesis (DOS) and the related biological trials focused primarily on expanding the current chemical space of paramount interest and 'fill in holes' wherein can be appropriately performed using Kohonen SOMs in simple and intuitively sensed manner. Practically, having novel structures in hands, virtual and/or real, as well as computational models listed above a specialist is then able to predict a wide

spectrum of core drug properties; moreover it can be performed in a simple and straightforward way. The calculated output then provides a comprehensive and convenient visual representation of multiparametric tangled data intricately hidden under the molecular descriptor covering and the related score values for the each object tested that can be reasonably used as key indicators for the target property prediction. Furthermore, the map constructed provides a convenient tool for the thorough analysis and elucidation of the knowledge obtained. For example, on the basis of the integral inspection of particular descriptor distribution within the map a specialist may make a reasonable suggestion towards the relations revealed.

To date, many novel variants and modifications of the basic Kohonen algorithm and learning rule are used in various applications including '*in silico*' drug design (for example, see [40]). A number of approaches are aimed at defining a better match of an input occurrence with the internal images, are currently applied. In addition, the activation neighborhood function that modulates the sensitivity of each Kohonen element can be defined in many ways. For example, such methods include [41]: Tree Structured SOM, Minimum Spanning Tree SOM, Neural Gas, Convex Combination, Duane Desieno Algorithm, Noise Technique, Growing Cell Structures, Two learning stages, 3D-architecture and so on. Several software packages that incorporate self-organizing Kohonen mapping are currently available from various sources. Some of them can be successfully applied in chemistry applications.

Nonlinear Sammon mapping

Nonlinear mapping is an advanced machine-learning technique for improved data mining and visualization. This method, originally introduced by Sammon [42], represents a multivariate statistical technique closely related to MDS described above. Just like MDS, the central objective of the Sammon approach is to approximate the local geometry and a common topological structure of multidimensional data on a visually interpretable two-dimensional or three-dimensional plot. The fundamental goal of this method is to substantially reduce the high-dimensionality of initial data set into the low-dimension feature space, irrespective of the number of dimensions from which it is constructed.

The classical Sammon algorithm attempts closely to approximate global geometric relationships observed across the whole space of input vector samples. The Sammon algorithm is arguably the most commonly used approach for accurate dimensionality reduction, but the main problem arising from this technique is that it does not scale well with the size of the input data set. Several attempts have recently been undertaken to reduce the complexity and difficulty of the task (e.g. [43]). The key advantage of the nonlinear technique over the Kohonen network is that it often provides much greater detail about individual compounds and the corresponding interrelationships as demonstrated by the following example. Target projection was carried out entirely using a set of 12-dimensional autocorrelation descriptors and the Euclidean metric was used as a pairwise measure of dissimilarity among the examined structures including xanthene, cubane and adamantane libraries [29]. The resulting 2D-plane is shown in Figure 4.

A wide number of different statistical problems can be effectively solved using the Sammon methodology. For example, a new

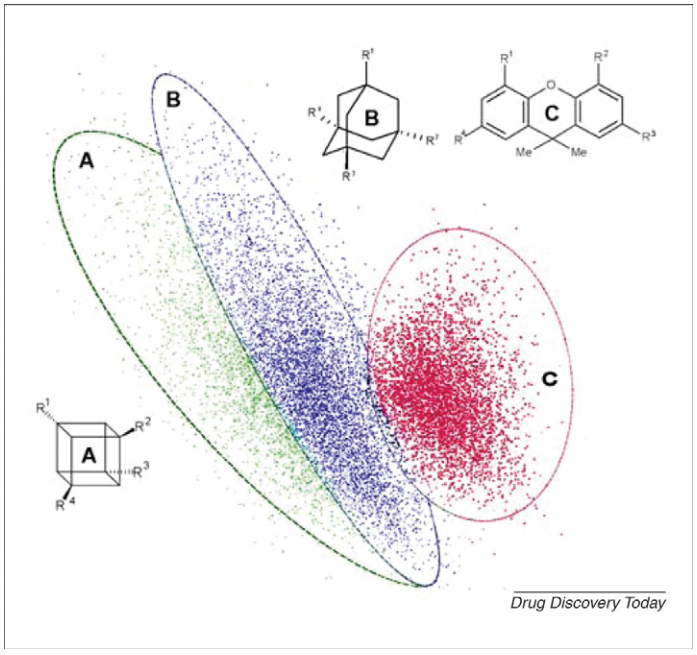

FIGURE 4

A nonlinear 2D-Sammon projection of the cubane (a), adamantane (b), and xanthene (c) combinatorial libraries. The feature space presented was constructed using a set of 12-dimensional autocorrelation descriptors as input variables and the Euclidean distances as a similarity metric.

method for analyzing protein sequences was also introduced by Agrafiotis [44] based on the Sammon algorithm. When applied to a family of homologous sequences, this method is able to capture the essential features of the similarity matrix, and provides a faithful representation of chemical or evolutionary distance in an intuitive way. The key merits of the new algorithm were clearly demonstrated using examples from the protein kinase family. The algorithm was also investigated as a means of visualizing and comparing large compound collections as well as ADME assessment, represented generally by a set of various molecular descriptors [36].

IsoMap

IsoMap is a promising new technique for resolving problems related to the MDS algorithm [45]. The key idea is to find a quasi-isometric, low-dimensional representation of a set of high-dimensional data points. By analogy with Sammon mapping, the algorithm attempts to preserve all pairwise geodesic (curvilinear) distances between the input data points within the whole feature space as closely as possible. This distance is defined narrowly as the shortest path between a pair of sample points and incorporated with the classical MDS. The algorithm provides a simple method for estimating the intrinsic geometry of a data manifold on the basis of a rough estimate of each datum point's neighbors on the manifold. It is highly efficient and generally applicable to a broad range of data sources and dimensionalities. It should be noted, however, that this technique produces favorable results only when a representative sampling of input data points across the manifold is presented.

Despite some limitations, IsoMap was successfully applied to visualization of different types of biomedical data [46].

As recently shown, a graphical visualization of IsoMap models provides a useful tool for exploratory analysis of protein microarray data sets. In most cases, IsoMap planes can adequately explain more of the variance presented in the microarray data in contrast to PCA or MDS [47]. In addition, for more detailed analysis of the large protein folding data sets (molecular dynamics trajectories of an SH3 protein model) the algorithm was subsequently modified using landmark points in the geodesic distance

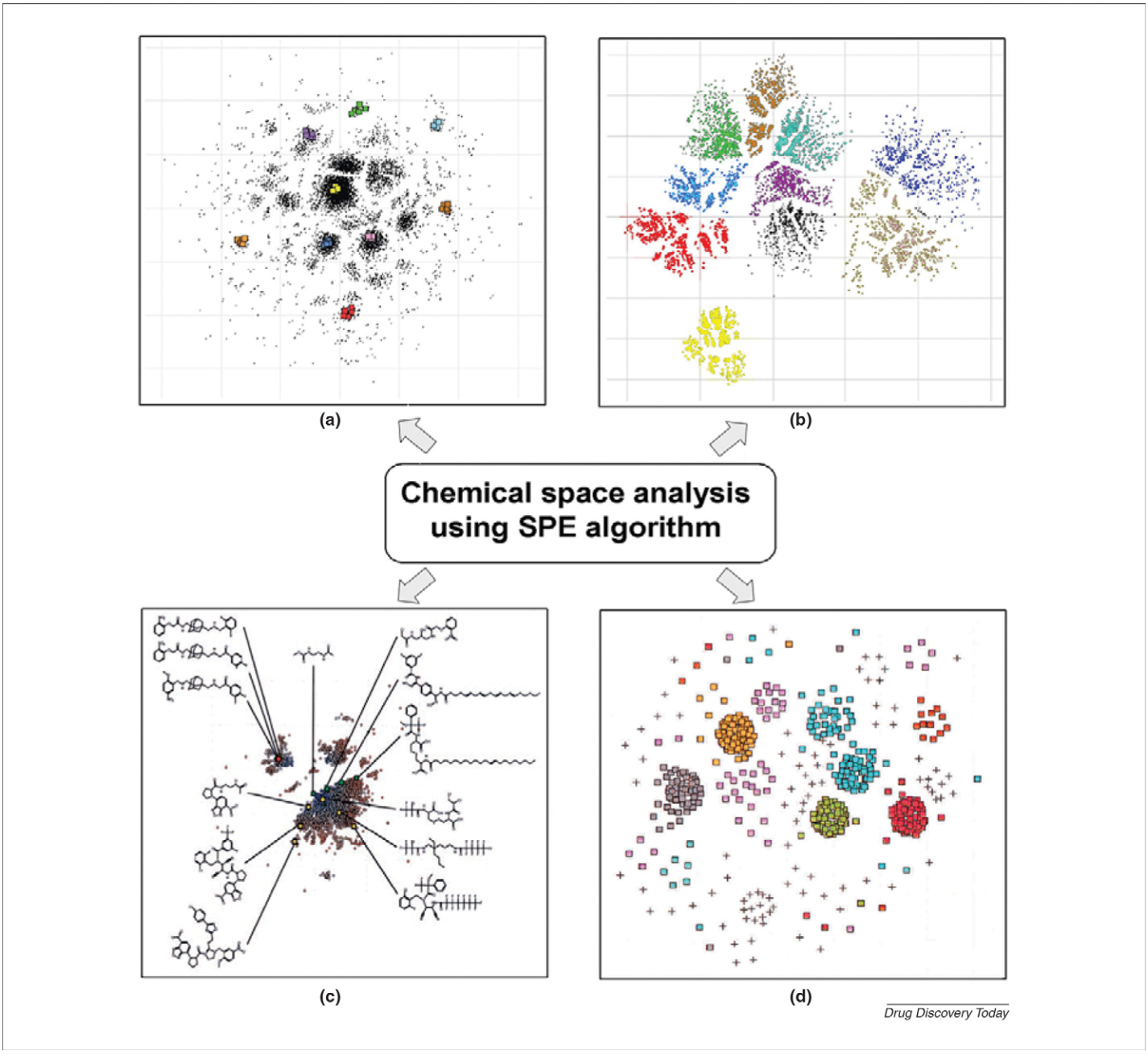

FIGURE 5

(a) Represents a two-component virtual combinatorial library [48] containing 10 000 compounds, derived by combining 100 amines and 100 aldehydes using the reductive amination reaction. Each of the products was described by 117 topological descriptors including molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev–Trinajstic indices, and topological state indices. To eliminate strong linear correlations, which are typical of graph-theoretic descriptors, the data were normalized and decorrelated using PCA. Molecular dissimilarity was defined as the Euclidean distance in the latent variable space formed by the 23 principal components that accounted for 99% of the total variance in the data. (b) Represents a four-component virtual combinatorial library containing 10 000 compounds derived by combining 10 carboxylic acids, 10 primary amines, 10 aldehydes and 10 isonitriles using the Ugi reaction. Each of the products was described by a 166-dimensional binary fingerprint, where each bit encoded the presence or absence of a particular structural feature in the target molecule, as defined in the ISIS chemical database management system. Molecular dissimilarity was calculated on the basis of the Tanimoto coefficient. The resulting maps exhibit the compact clusters that correspond to distinct chemical classes of the tested structures resulting from the discrete nature of the descriptors and the diversity of the chemical fragments employed. These maps that are discussed in greater detail in [49] are illustrative of the ability of SPE to identify natural clusters in the data set without prior knowledge or expert guidance. (c) Two-dimensional stochastic proximity map of different types of organic compounds. Therefore, this algorithm can be easily applied to other types of binary fingerprints that typically consist of a few thousand bits or various high-dimensional descriptor spaces that are commonly employed in different QSAR studies [50]. (d) Two-dimensional SPE maps of the Manning kinase domains using a neighborhood radius of 0.89.

calculations. On the basis of the results obtained, it was shown that for accurate interpretation and analysis of the original data the approach was far more effective versus the linear techniques of dimensionality reduction, such as PCA. Whereas the Euclidean metric is based directly on quadratic relationships, IsoMap scales with the third power of the number of data points and then becomes computationally prohibitive for processing and visualizing large data sets. As mentioned in the next section, a similar scaling problem also limits the SPE algorithm, a related approach that produces globally ordered maps by constructing locally linear relationships among the input data points.

Stochastic proximity embedding (SPE)

Another promising self-organizing algorithm, SPE, is a technique recently developed by Agrafiotis *et al.* [48] for embedding a set of related observations into a low-dimensional space that preserves the intrinsic dimensionality and metric structure of the data. It partly addresses the key limitations of IsoMap and Sammon methodology. First, it circumvents the calculation of estimated geodesic distances between the embedded objects and, secondly, it uses a pairwise refinement strategy that does not require the complete distance or proximity matrix maintaining a minimum separation between distant objects and scales linearly with the size of the data set. In other words, SPE utilizes the fact that the geodesic distance is always greater than or equal to the input proximity if the latter is an accurate metric. Unlike previous stochastic approaches of nonlinear manifold learning that preferentially preserve local over global distances, the method operates by estimating the proximities between remote objects as lower bounds of their true geodesic distances and uses them as a means to impose global structure. Thus, it can reveal the underlying geometry of the manifold without intensive nearest neighbor or shortest path computations. Therefore, it can preserve the local geometry and the global topology of the manifold better than previous approaches and it can be effectively applied to very large data sets that are intractable by conventional embedding procedures.

One potential limitation of SPE is related to numerous adjustable and internal parameters. Thus, just like IsoMap, SPE strongly depends on the choice of the neighborhood radius. If it is too large, the local neighborhoods will include data points from other branches of the manifold, shortcutting them and leading to substantial errors in the final embedding, whereas if it is too small, it may lead to discontinuities, causing the manifold to fragment into a large number of disconnected clusters.

SPE can be effectively applied in different classification tasks and the method has been successfully used for analysis of the 'Swiss roll' data set [48–50]. SPE can also produce meaningful low-dimensional representations of more complex data sets that do not have a clear manifold geometry. Thus, the embedding of a combinatorial chemical library, illustrated in Figure 5, shows that SPE is able to preserve local neighborhoods of closely related compounds while maintaining a chemically meaningful global structure. For example, amination and Ugi virtual combinatorial libraries have been recently analyzed using the modified SPE algorithm (Figure 5a,b).

SPE has been applied to an important class of distance geometry problems including conformational analysis [51], NMR structure determination, and protein structure prediction [52]. For example, a successful implementation of stochastic proximity embedding in combination with the self-organizing superimposition (SOS) algorithm is a promising computational approach to conformational sampling and conformational search that was recently described [53]. SPE has also been applied to classify and visualize protein sequences as well as to reduce the intrinsic dimensionality and metric structure of the data obtained from genomic and proteomic research (Figure 5d) [54]. The effectiveness of the algorithm can also be illustrated using examples from the protein kinase and nuclear hormone receptor superfamilies [44]. Using the progressive multiple sequence alignment technique the method has produced informative maps that preserved the intrinsic structure and clustering of the input data. Finally, Demartines and Hérault [55] have recently developed a similar computational algorithm for nonlinear dimensionality reduction currently known as curvilinear component analysis (CCA).

Conclusion

This article presents a mini review specifically focused on advanced computational mapping techniques currently applied to '*in silico*' drug design and development. Several novel mapping approaches to analysis of the structure–activity relationships within the chemical space of different types and structure were comparatively discussed in order to provide a better understanding of fundamental principles of dimensionality reduction. Owing to space limitations, many promising mapping techniques and their applications remain beyond the scope of this review, the techniques described therefore highlight the role mapping algorithms currently play in computational chemistry and drug discovery. We believe this area will probably expand in the future owing to the increasing amounts of data being deposited in public databases as well as internally generated in pharmaceutical companies.

References

1 Howe, T.J. *et al.* (2007) Data reduction and representation in drug discovery. *Drug Discov. Today* 12, 45–53

2 Engels, M.F.M. and Reijmers, T.H. (2004) Data mining application in drug discovery. In *Computational Medicinal Chemistry for Drug Discovery*, (794) (Bultinck, P., Tollenaere, J.P., Langenaeker, W., De Winter, H., eds) CRC Press

3 Downs, G.M. and Barnard, J.M. (2002) Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry*, (vol. 18) (Lipkowitz, K.B. and Boyd, D.B., eds) pp. 1–40, VCH

4 Suykens, J.A.K. (2003) Data visualization and dimensionality reduction using kernel maps with a reference point. In *Internal Report 07-22, ESAT-SISTA*. Leuven

5 Sha, F. and Saul, L.K. (2005) Analysis and extension of spectral methods for nonlinear dimensionality reduction. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 119. pp. 785–792

6 Faloutsos, C. and Lin, K.-I. (1995) FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of 1995 ACM SIGMOD, SIGMOD RECORD, June 24* pp. 163–174

7 Johnson, M.A. and Maggiora, G.M. (1990) *Concepts and Applications of Molecular Similarity*. Wiley

8 Jain, A.K. *et al.* (1999) Data clustering: a review. *ACM Comput. Surv.* 31, 264–323

9 Herman, I. (2000) Graph visualization and navigation in information visualization: a survey. *IEEE Trans. Vis. Comp. Graph.* 6, 24–43

10 Strehl, A. and Ghosh, J. (2003) Relationship-based clustering and visualization for high-dimensional data mining. *INFORMS J. Comput.* 15, 208–230
11 Agrafiotis, D.K. *et al.* (2007) Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.* 47, 69–75
12 Agrafiotis, D.K. *et al.* (2001) Multidimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* 22, 488–500
13 Hassan, M. *et al.* (1996) Optimization and visualization of molecular diversity of combinatorial libraries. *Mol. Div.* 2, 64–74
14 Taylor, R. (1995) Simulation analysis of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. *J. Chem. Inf. Comput. Sci.* 35, 59–67
15 Boyd, S.M. *et al.* (1995) Characterizing the geometrical diversity of functional groups in chemical databases. *J. Comput. Aid. Mol. Des.* 9, 417–424
16 Mount, J. *et al.* (1996) *IBC 6-th Annual Conference on Rational Drug Design, December 11-12*, Coronado, USA
17 Goulon, A. *et al.* (2006) Graph machines and their applications to computer-aided drug design: a new approach to learning from structured data. In *Lecture Notes in Computer Science*. Springer pp. 1–19
18 Nielsen, T.O. *et al.* (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* 359, 1301–1307
19 Nadler, B. *et al.* (2006) Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems. *Appl. Comput. Harm. Anal.* 21, 113–127
20 Zhang, Z. and Zha, H. (2003) Local linear smoothing for nonlinear manifold learning.. In *Technical Report, CSE-03-003*. Department of Computer Science and Engineering, Pennsylvania State University, University Park
21 Willshaw, D.J. and von der Malsburg, C. (1976) How patterned neural connections can be set by self-organization. *Proc. R. Soc. Lond. Ser. B* 194, 431–445
22 Kohonen, T. (1988) *Self-organization and Associative Memory* (3rd ed.), Spingler-Verlag
23 Linde, Y. *et al.* (1980) An algorithm for vector quantizer design. *IEEE Transac. Commun.* 28, 84–95
24 Kohonen, T. (1993) Physiological interpretation of the self-organizing map algorithm. *Neural Netw.* 6, 895–905
25 Gasteiger, J. *et al.* (1994) The beauty of molecular surfaces as revealed by self-organizing neural networks. *J. Mol. Graph.* 12, 90–97
26 Balakin, K.V. *et al.* (2004) In silico estimation of DMSO solubility of organic compounds for bioscreening. *J. Biomol. Screen.* 9, 22–31
27 Korolev, D. *et al.* (2003) Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.* 46, 3631–3643
28 Schneider, G. and Wrede, P. (1998) Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* 70, 175–222
29 Sadowski, J. *et al.* (1995) Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem. Int. Ed. Engl.* 34, 2674–2677
30 Schneider, G. *et al.* (1995) A peptide selection scheme for systematic evolutionary design and construction of synthetic peptide libraries. *Minim. Invas. Med.* 6, 106–115
31 Bauknecht, H. *et al.* (1996) Locating biologically active compounds in medium-sized heterogeneous datasets by topological auto-correlation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* 36, 1205–1213
32 Anzali, S. *et al.* (1997) Endothelin antagonists: search for surrogates of methylendiox–yphenyl by means of a Kohonen neural network. *Bioorg. Med. Chem. Lett.* 8, 11–16
33 Polanski, J. (1997) The receptor-like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.* 37, 553–561
34 Balakin, K.V. *et al.* (2004) Quantitative structure–metabolism relationship modeling of the metabolic *N*-dealkylation rates. *Drug Metab. Dispos.* 32, 1111–1120
35 Balakin, K.V. *et al.* (2004) Kohonen maps for prediction of binding to human cytochrome P450 3A4. *Drug Metab. Dispos.* 32, 1183–1189
36 Balakin, K.V. *et al.* (2005) Comprehensive computational assessment of ADME properties using mapping techniques. *Curr. Drug Discov. Technol.* 2, 99–113
37 Ekins, S. *et al.* (2006) Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning, Kohonen and Sammon mapping techniques. *J. Med. Chem.* 49, 5059–5071
38 Aronov, A.M. *et al.* (2007) Applications of QSAR methods to ion channels. In *Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals* (Ekins, S., ed.), pp. 353–389, John Wiley and Sons
39 Chekmarev, D.S. *et al.* (2008) Shape signatures: new descriptors for predicting cardiotoxicity *in silico*. *Chem. Res. Toxicol.* 21, 1304–1314
40 Kohonen, T. (1996) *Self-organizing Maps*. Springer-Verlag
41 Fritzke, B. (1994) Growing cell structures—a self-organizing neural network for unsupervised and supervised learning. *Neural Netw.* 7, 1441–1460
42 Sammon, J.W. (1969) A non-linear mapping for data structure analysis. *IEEE Trans. Comp.* C-18, 401–409
43 Agrafiotis, D.K. and Lobanov, V.S. (2000) Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* 40, 1356–1362
44 Agrafiotis, D.K. (1997) A new method for analyzing protein sequence relationships based on Sammon maps. *Prot. Sci.* 6, 287–293
45 Tenenbaum, J.B. (1998) Mapping a manifold of perceptual observations. The MIT Press pp. 682–688
46 Lim, I.S. *et al.* (2003) Planar arrangement of high-dimensional biomedical data sets by Isomap coordinates. *Proc. of the 16th IEEE Symposium on Computer-Based Medical Systems*, Mount Sinai Medical School, New York. pp. 50–55
47 Dawson, K. *et al.* (2005) Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using Isomap, a nonlinear algorithm. *BMC Bioinform.* 6, 1–17
48 Agrafiotis, D.K. and Xu, H. (2002) A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci.* 99, 15869–15872
49 Agrafiotis, D.K. and Xu, H. (2003) A geodesic framework for analyzing. molecular similarities. *J. Chem. Inf. Comput. Sci.* 43, 475–484
50 Rassokhin, D.N. and Agrafiotis, D.K. (2003) A modified update rule for stochastic proximity embedding. *J. Mol. Graph. Model.* 22, 133–140
51 Spellmeyer, D.C. *et al.* (1997) Conformational analysis using distance geometry methods. *J. Mol. Graph. Model.* 15, 18–36
52 Havel, T.F. and Wuthrich, K. (1985) An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. *J. Mol. Biol.* 182, 281–294
53 Zhu, F. and Agrafiotis, D.K. (2007) Self-organizing superimposition algorithm for conformational sampling. *J. Comput. Chem.* 28, 1234–1239
54 Farnum, M.A. *et al.* (2003) Exploring the nonlinear geometry of protein homology. *Prot. Sci.* 12, 1604–1612
55 Demartines, P. and Hérault, J. (1997) Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. Neural Netw.* 8, 148–154