Chemography of Natural Product Space

Tomoyuki Miyao1,2*, Daniel Reker2*, Petra Schneider2, Kimito Funatsu1, Gisbert Schneider2

1 Department of Chemical Systems Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan
2 Department of Chemistry and Applied Biosciences, Institute of Pharmaceutical Sciences, ETH Zurich, Zurich, Switzerland

Abstract

We present the application of the generative topographic map algorithm to visualize the chemical space populated by natural products and synthetic drugs. Generative topographic maps may be used for nonlinear dimensionality reduction and probabilistic modeling. For compound mapping, we represented the molecules by two-dimensional pharmacophore features (chemically advanced template search descriptor). The results obtained suggest a close resemblance of synthetic drugs with natural products in terms of their pharmacophore features, despite pronounced differences in chemical structure. Generative topographic map-based cluster analysis revealed both known and new potential activities of natural products and drug-like compounds. We conclude that the generative topographic map method is suitable for inferring functional similarities between these two classes of compounds and predicting macromolecular targets of natural products.

Abbreviations

CATS: chemically advanced template search
COBRA: Collection of Bioactive Reference Analogs
DI: deoxydihydroisoflindisso1
DNP: Dictionary of Natural Products
EM: expectation-maximization
GMM: Gaussian mixture model
GPCR: G-protein coupled receptor
GTM: generative topographic map
5-HT: 5-hydroxytryptamine
MACCS: molecular access system
MOE: Molecular Operating Environment
NCE: new chemical entity
PCA: principal component analysis
RBF: radial basis function
RMSE: root mean square error
SOM: self-organizing map
WOMBAT: World of Molecular Bioactivity

Introduction

Natural products have a long-standing history as a source for innovative compounds in drug discovery [1–3]. A rationale for their success is the historic evolutionary exploration of chemical modifications leading to compounds containing privileged structural motifs with biophoric properties [4,5]. It has been estimated that for more than half of the published NCEs for therapeutic use, natural products served as an inspiration [6], with a particular emphasis on anticancer agents [7]. First studies have reported computer-assisted natural product deorphaning and also presented algorithmic advances for automated structure elucidation [8–11], suggesting that pharmacophore models derived from bioactive natural products may guide the development of drug-like, chemically tractable natural product mimetics. Here, we show how a global perspective on the pharmacophores found in natural products and synthetic drugs may be exploited for drug discovery by visualizing landscapes of pharmacophoric traits. These give insightful hints for relating natural products to synthetic compounds and assessing their potential polypharmacology. The approach also provides a means for identifying sparsely populated biophoric regions of chemical space.

Chemography is an umbrella term describing computational methods for the visual inspection of typically two-dimensional representations of...
Motivated by the fact that pharmacophore representations allow for a meaningful mixing of natural products and synthetic drugs when mapped with the help of SOMs, here we explored GTMs for a meaningful mixing of natural products and synthetic drugs. Motivated by the fact that pharmacophore representations allow for projecting a data instance (here: a molecule) not only to one point on the map but instead calculates the activation for every single Gaussian, so that we obtain a fuzzy location of every compound on the map (Fig. 1). The quality of the map can be assessed through calculating the RMSE (Eq. 1) of back-projected data points.

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \| y(i, W) - x_i \|^2}
\]

where \(N\) is the number of compounds, \(i\) is the position of compound \(i\) on the map (also referred to as the latent space), \(x_i\) is the descriptor vector of compound \(i\), and \(y\) is the back-projection function from the latent space to the original data space. In simple terms, Eq. 1 corresponds to the loss of information by projecting from the high-dimensional original data space spanned by the chemical descriptors to the lower-dimensional latent space.

The number of RBFs, the variance (width) of the basis functions, the regularization parameters, and the number of latent points (map size) govern GTM training. The parameters related to the RBFs control the linearity (smoothness) of the map, and the regularization parameter helps avoid overfitting. The number of Gaussian distribution functions in the high-dimensional original data space gives the number of latent points in the projection. Each latent point is connected to the mean value of a Gaussian distribution in the original data space, and the sum of the Gaussian distributions captures the underlying data distribution.

We generated a GTM for a total of 157929 natural products and a small but carefully curated collection of 12644 drug-like synthetic compounds for which macromolecular targets are annotated (COBRA [30]). The molecules were represented by 210-dimensional CATS (version 2) pharmacophore descriptors [31, 32]. After GTM training, the RMSE (Eq. 1) for all compounds in both data collections had a value of 0.99. For the RMSE calculation,
every compound was projected at the posterior mean on the map. A higher number of RBFs resulted in a better fit to the original data, but at the same time increased the complexity of the model (Fig. 2). We decided that 100 RBFs represented an acceptable compromise to achieve a generalization of the compound distribution while still leading to an appropriate quality of the mapping. Fig. 2 shows that our choice was reasonable in terms of the change of the RMSE against the number of RBFs.

To identify positions (clusters) on the GTM that are populated with natural products and drug-like molecules, we projected each of the compounds onto exactly one point on the map using the maximum posterior probability criterion. In agreement with an earlier study using SOM projections [20], the natural products and drug-like compounds intermixed strongly (Fig. 3). In fact, more than 60% (856 of 1424) of the natural product clusters also contained synthetic drugs. Importantly, these co-clustered regions contained more than 70% (110 788) of all the natural products. Only 10 drug-like compounds were projected to a total of six clusters that were not occupied by any natural product. While we observed mostly intermixed clusters of the COBRA drug set and natural products sharing pharmacophoric traits, natural products exclusively populated 568 other clusters. These clusters might point to pharmacophores that have only rarely been explored for synthetic drug design. In an attempt to further analyze this observation and clarify whether it is caused by the biased sample size, i.e., the small set of synthetic drugs compared to the much larger DNP set, we projected the ChEMBL compound collection [33] onto the map (the large number of ChEMBL entries prevented GTM retraining). We considered only 98.8% of the ChEMBL data (n = 1 351 370 without duplicate DNP entries) by avoiding projecting compounds that clearly reside outside of the applicability domain. We defined the model’s applicability domain as the region in which the probability density of a molecule was larger than a threshold value corresponding to the 99.5 percentile of the training data (Eq. 3). The ChEMBL compounds spread almost over the complete spanned chemical product space, including many of the clusters not populated by COBRA compounds. It might therefore be a worthwhile exercise to systematically analyze the ChEMBL entries for activity annotations and hypothesize related activities for the co-located natural products (ongoing). Nevertheless, it is important to keep in mind that there are many natural products and derivatives in ChEMBL which are not part of the DNP (Fig. 4), and that the ChEMBL data were forced on the GTM trained only with DNP and COBRA. We consequently did not consider ChEMBL data for subsequent GTM analysis. This decision is supported by the statistically insignificant difference of compound properties between COBRA and ChEMBL (Table 1), which motivates the use of small, curated compound sets as surrogates for much larger collections.

We investigated representative natural products located in regions not populated by COBRA drugs (Fig. 3A, B). In region A, we observed highly hydrophilic compounds [e.g., triaspidin (1, Fig. 5) in cluster (29/9)]. While this may be an undesired property in many drug discovery projects, pronounced lipophilicity is not necessarily a requirement for natural products as their secretion and uptake might be governed by different mechanisms or they might act intracellularly. GTM region B is populated by guaianolide derivatives [e.g., vestenolide (2) in cluster (7/13)]. This natural product scaffold has recently gained attention for multiple indications but is still only scarcely studied [34, 35]. These select examples reveal the potential of the map to suggest

Fig. 2 Mapping error of the GTM projection (RMSE, Eq. 1) with respect to the number of RBFs used to define the map’s resolution. The error bars present the standard deviations computed from 30 experiments with randomly initialized mapping parameters (W, b). The curve gives the RMSE of models initialized through PCA analysis of the original data.

Fig. 3 GTM projection of pharmacophore space. The map shows 99.5% of the training data lacking extreme outliers. A magenta cross marks clusters that contain natural products (157 265 compounds from DNP), red open squares mark clusters that contain drugs or leads from COBRA (12 455 compounds), and orange dots show the location of the projected ChEMBL library (1 351 370 compounds). Note that ChEMBL data were not used for GTM training. Background shading reflects the position of the cluster centroids according to aromaticity, indicating the average aromaticity of the clustered compounds. The coloring ranges from white (no compound aromatic) to black (all compounds aromatic). Upper-case labels (A)–(D) highlight areas populated by natural products and drugs from both ChEMBL and COBRA (cf. Fig. 5), while lower-case labels (a)–(e) point to regions of chemical space containing natural products and ChEMBL compounds only (cf. Fig. 4). (Color figure available online only.)
new routes for filling pharmacophoric holes in synthetic compound libraries. We also compared natural products and drug-like compounds that are actually co-clustered. This analysis allowed us to assess whether functionally similar compounds were grouped together. GTM region C consists of prostaglandin derivatives from both the natural product collection [e.g., prostaglandin J2 (3) at position (9/35)] and the COBRA drug database [e.g., gemeprost (4) in cluster (8/34)]. Region D also features compounds that possess convincingly similar pharmacophore patterns in spite of apparent differences in their chemical structures. Two representative examples are the natural coumarin derivative mexoticin (5), which has been shown to inhibit platelet aggregation [36], and synthetic compound 6 [cluster (26/26)]. The coumarin-derived scaffold of mexoticin (5) contains the pharmacophore of psoralen, which is known to induce insomnia as a side effect in patients [37]. The co-clustered compound 6 is a synthetic orexin receptor antagonist that was developed as a treatment for sleep disorders [38]. Despite their apparent difference in chemical structure (MACCS-key based structural Tanimoto index = 0.27), the GTM suggests that the two compounds share a pharmacophore pattern. These observations motivate the testing of mexoticin for orexin receptor binding, and compound 6 for effects on platelet aggregation. In this way, the GTM may be used for predicting the macromolecular targets of natural products, in analogy to the SOM [11, 39] and property-based methods [22].

### Table 1

<table>
<thead>
<tr>
<th>Data set (unique compounds)</th>
<th>H-bond acceptors</th>
<th>H-bond donors</th>
<th>Rotatable bonds</th>
<th>Rings</th>
<th>SlogP</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNP (n = 157 929)</td>
<td>5.9 ± 5.8</td>
<td>2.9 ± 3.6</td>
<td>6.6 ± 6.9</td>
<td>3.5 ± 2.4</td>
<td>2.5 ± 3.3</td>
<td>447 ± 256</td>
</tr>
<tr>
<td>COBRA (n = 12 644)</td>
<td>3.3 ± 2.0</td>
<td>1.5 ± 1.5*</td>
<td>7.4 ± 5.1*</td>
<td>3.4 ± 1.4*</td>
<td>2.7 ± 2.7</td>
<td>417 ± 136*</td>
</tr>
<tr>
<td>ChEMBL (n = 1 368 655)</td>
<td>3.6 ± 3.5</td>
<td>1.5 ± 2.6*</td>
<td>7.7 ± 8.9*</td>
<td>3.4 ± 1.6*</td>
<td>2.8 ± 2.9</td>
<td>423 ± 245*</td>
</tr>
</tbody>
</table>

* No significant difference of the means (Welch’s two-sided t-test, p value > 10^−6)
We investigated whether the GTM appropriately spans the chemical space of synthetic COBRA compounds and natural products, and analyzed representative structures at regions that are heavily populated. We observed pronounced structural differences in these clusters. For example, a group of carbohydrate-containing natural products is represented by quercetin diglycoside (7) located at position (40/13) on the map. We found 11-oxooctacosanal (8) at position (11/40) representing a cluster of fatty acids and derivatives. In contrast to the natural products, the cluster representatives of synthetic drugs are not as easily structurally distinguishable. However, they represent ligands for different classes of biomolecules, for example GPCRs (e.g., compound 9) or nuclear receptors (e.g., compound 10). We aim at connecting the structural separation of natural products and target-specific regions on the map for the identification of meaningful relationships of natural product structures with synthetic ligands of pharmacologically relevant biomolecules. For example, we found DI [40] [position (1/32)] as a representative natural product in the region populated by vitamin D receptor ligands. The pharmacophore of DI closely resembles the one of vitamin D itself but constitutes a scaffold-hop from the secosteroids to the closed steroidal form. While this structural modification does not change the relative positioning of the pharmacophores, as correctly recognized by the CATS descriptor, it locks the hydrogen bond donor function in the 6-s-cis conformation [41]. It has been suggested that the 6-s-cis conformation is associated with the immediate response of vitamin D [42]. Further analysis of DI might reveal its role in different cellular processes and help to explain these effects. Chemographic methods can help in formulating motivated hypotheses for these experiments.
Ligands of serotonin (5-HT) receptors occur in six distinctive clusters on different positions on the map (Fig. 6). Interestingly, five of these are adjacent or intermixed with ligands of the adrenergic receptor. This is in line with their known pharmacological cross-activity [43]. However, one of the clusters is completely disconnected from any area containing adrenergic receptor ligands [around position (32/12)]. We investigated this area further and in fact found several compounds known to be selective for certain 5-HT receptor subtypes without activity on the adrenergic receptor family, for example alsetron [position (32/12)] [44], eplivasin [position (33/11)] [45], and ramelteon [position (34/11)] [44]. Apparently, the resolution of the map is high enough to distinguish such subtle changes in pharmacophores.

GTMs are gaining increasing attention in cheminformatics [25, 46]. In natural product-related studies, GTMs have been applied to applications outside of drug discovery, for example, in the analysis of the content of fish oil extracts [47]. An exception is the study by Owen et al. that used structural MACCs key fingerprints to distinguish drugs, combinatorial synthetic compounds, and natural products [48]. Here, we have introduced GTMs as a technique for dimensionality reduction and target prediction for natural products based on molecular pharmacophore representations. We show that this concept allows for relating natural products and synthetic compounds in spite of clearly observable structural differences, fully in line with results previously acquired with SOMs [11, 20]. Results suggest that the resolution of a GTM is sufficient to identify functional relationships between natural products and synthetic drugs. The concept of analyzing regions of natural product space for a lack of synthetic compounds or the presence of compounds with a desired polypharmacological profile will make such chemographic methods a helpful tool for natural product-inspired drug discovery.

Materials and Methods

Data

For GTM analysis, we compiled the natural products contained in the Chapman & Hall/CRC Dictionary of Natural Products (DNP v20.1 DNP, 210273 compounds; http://dnp.chemnetbase.com) [49]. Drugs and drug-like bioactive compounds were taken from COBRA (v12.6, 13,702 compounds; inSili.com LLC) [30]. ChEMBL compound data were compiled from database version 19 (https://www.ebi.ac.uk/chembl/) [33]. We removed all duplicates (5088 compounds) and structures also present in the DNP (31009 compounds) from the ChEMBL collection, which resulted in 1368655 remaining ChEMBL entries. All molecules were preprocessed with the MOE wash node (v2011.10, Chemical Computing Group) as implemented in KNIME v2.9.4 [50] using the options “protonate strong bases”, “deprotonate strong acids”, “remove minor components”, “disconnect salts”, and “remove lone pairs”. Duplicate structures were removed by grouping according to canonical SMILES representations. This procedure resulted in 157929 compounds from the DNP and 12644 compounds from the COBRA database. All molecules were described in terms of pharmacophore patterns using our in-house CAS2 descriptor implementation with a correlation distance of 0–9 bonds and type-sensitive scaling [31]. Consequently, each molecule was represented by a 210-dimensional topological pharmacophore representation.

Generative topographic map

In GTMs, every latent point is connected to the point in the original data space according to Eq. 2. Simultaneously, the projected point is the mean value of a Gaussian and the sum of these Gaussians describes the data distribution (Eq. 3).

\[
y(l, W) = W \Phi(l)
\]

\[
p(x|l, W, \beta) = \frac{1}{K} \sum_{i=1}^{K} \frac{\beta}{2\pi} \exp \left\{ -\frac{\beta}{2} \| x - y(l, W) \|^2 \right\}
\]

In Eq. 2, \(l\) is a latent point, \(K\) is the number of latent points, and \(\Phi\) is an \(M\)-dimensional vector consisting of RBFs evaluated at \(l\). The matrix parameter \(W (D \times M)\) governs the projection from a point in the latent space to the point in the original data space, where \(D\) is the dimensionality (descriptor vector cardinality) of the original data space. In Eq. 3, the sum of Gaussian distributions gives the probability distribution in the data space. Each Gaussian has the mean value \(y(l, W)\) with variance \(\beta^{-1}\). By using the EM algorithm, locally optimized parameters \((W, \beta)\) were obtained. We trained a GTM with \((40 \times 40) = 1600\) latent points that served as Gaussian cluster centers. The maps were constructed using the GTM toolbox v1.0 [51] from Matlab 2014a (The MathWorks, Inc.). During the training, we used 100 RBFs that were aligned on the lattice in latent space. The width of the RBFs was set to the distance between neighboring RBF centers. The regularization parameter was set to 0.001. We initialized the map through PCA of the original data. Note that for the estimation of the standard deviation of the mapping error (RMSE, Eq. 1), we randomly initialized the mapping parameters \((W, \beta)\). Applicability domain was determined based on Eq. 3 with optimized parameters. We set a 99.5 percentile of the values of descending ordered training data as the threshold value.

Acknowledgements

ETH Zürich and the OPO Foundation Zurich funded this research study.

Conflict of Interest

G.S. and P.S. are cofounders of inSili.com LLC, Zürich, and act as consultants in the pharmaceutical and chemical industry.

References
