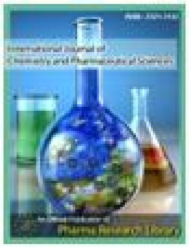
International Journal of Chemistry and Pharmaceutical Sciences

Journal Home Page:www.pharmaresearchlibrary.com/ijcps



REVIEW ARTICLE

Pharmacophore Modeling: Virtual Screening and Pharmacophore Identification in Drug Discovery

G. Mahalakshmi*[1], J. Amutha Iswarya Devi[1], S.R. Senthilkumar[2], N. Venkateshan[1]

*[*1]Department of Pharmaceutical Chemistry, Arulmigu Kalasalingam college of Pharmacy, Anand Nagar, Krishnankoil-626126, Tamil Nadu, India.*
[2]Department of Pharmaceutics, Arulmigu Kalasalingam College of Pharmacy, Anand Nagar, Krishnankoil-626126, Tamil Nadu, India.

ABSTRACT

Pharmacophore modeling is a successful yet very diverse subfield of computer-aided drug design. The concept of the pharmacophore has been widely applied to the rational design of novel drugs. In this paper, we review the computational implementation of this concept and its common usage in the drug discovery process. Pharmacophores can be used to represent and identify molecules on a 2D or 3D level by schematically depicting the key elements of molecular recognition. The most common application of pharmacophores is virtual screening, and different strategies are possible depending on the prior knowledge. However, the pharmacophore concept is also useful for ADME-tox modeling, side effect, and off-target prediction as well as target identification. Furthermore, pharmacophores are often combined with molecular docking simulations to improve virtual screening. We conclude this review by summarizing the new areas where significant progress may be expected through the application of pharmacophore modeling; these include protein–protein interaction inhibitors and protein design.

Keywords: drug design, Docking studies, Molecular modeling, Pharmacophore.

ARTICLE INFO

Corresponding Author
G. Mahalakshmi
Department of Pharmaceutical Chemistry,
Arulmigu Kalasalingam college of Pharmacy,
Krishnankoil-626126, Tamil Nadu, India.
MS-ID: IJCPS3946

PAPER-QRCODE

ARTICLEHISTORY: Received 11 February 2019, **Accepted** 18 March 2019, **Available Online** 27 April 2019

Citation: G. Mahalakshmi, *et al. Pharmacophore Modeling: Virtual Screening and Pharmacophore Identification in Drug Discovery*. *Int. J. Chem, Pharm, Sci.*, 2019, 7(4): 96-104.

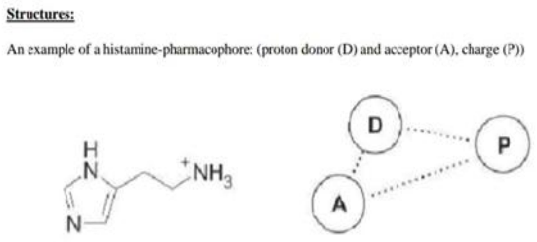
CONTENTS

1. Introduction

Definition: A Pharmacophore is an ensemble of steric, electrostatic and hydrophobic properties which is essential for optimal supramolecular interactions with a biolocial receptor, to modulate or inhibit a biological effect. A Pharmacophore does not represent a concrete molecule, but an abstract concept which describes the common molecular properties of interaction with the receptor. The Pharmacophore anchors the agent with the receptor. With pharmacophoric models one can define special properties (pharmacophoric points) based on the structure of the receptor or based on the structure of a known agent. This pharmacophoric points can be checked against a database of pharmacophores.

Characterization:

1) Location of the functional groups (e.g. proton donor/acceptor, hydrophobic parts)
2) Stabilization of the most effective conformation
3) Lipinski's rule of five: The following properties are essential for good permeation
    - The molecule has less than five proton-donators
    - The molecular weight is smaller than 500 Dalton
    - log P smaller than 5
    - The molecule has less acceptors than 10
    - The molecule should use biological transporters otherwise the ligand is attached too strong or it cannot be transported.
4) Minimum of pharmacophoric points:3



Structures:

An example of a histamine-pharmacophore: (proton donor (D) and acceptor (A), charge (P))

Figure 1

IUPAC defines a pharmacophore to be "an ensemble of steric and electronic features to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response. "

What is a pharmacophore?

Historical perspective: The original concept of the pharmacophore was developed by Paul Ehrlich during the late 1800s.43 At that time, the understanding was that certain "chemical groups" or func- tions in a molecule were responsible for a biological effect, and molecules with similar effect had similar functions in common. The word pharmacophore was coined much later, by Schueler in his 1960 book Chemobiodynamics and Drug Design, and was defined as activity"a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological."44 The definition of a pharmacophore was therefore no longer concerned with "chemical groups" but "patterns of abstract features." Since 1997, a pharmacophore has been defined by the International Union of Pure and Applied Chemistry as: "A pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response."

The pharmacophore should be considered as the largest common denominator of the molecular interaction features shared by a set of active molecules. Thus a pharmacophore does not represent a real molecule or a set of chemical groups, but is an abstract concept. Despite this clear definition, the term pharmacophore is often misused by many in medicinal chemistry to describe simple yet essential chemical functionalities in a molecule (such as guanidine or sulfonamides), or common chemical scaffolds (such as flavones or prostaglandins). Often the long definition is simplified to "A pharmacophore is the pattern of features of a molecule that is responsible for a biological effect," which captures the essential notion that a pharmacophore is built from features rather than defined chemical groups.

2. Pharmacophore concepts in CADD

While the pharmacophore concept predates any form of electronic computer, it has nevertheless become an important tool in CADD. Every type of atom or group in a molecule that exhibits certain properties related to molecular recognition can be reduced to a pharmacophore feature. These molecular patterns can be labeled as hydrogen bond donors or acceptors, cationic, anionic, aromatic, or hydrophobic, and any possible combinations. Different molecules can be compared at the pharmacophore level; this usage is often described as "pharmacophore fingerprints." When only a few pharmacophore features are considered in a 3D model the pharmacophore is sometimes described as a "query."

Pharmacophore fingerprint

While molecules are 3D entities, the pharmacophore representation reduces a molecule to a collection of features at the 2D or 3D level.47,48 A pharmacophore fingerprint is an extension of this concept, and typically annotates a molecule as a unique data string. All possible three-point or four-point sets of pharmacophore features (points) are enumerated for each ligand. The distance between the feature points is counted in bonds (for topological fingerprints), or by distance-binning when using 3D fingerprints (Figure 1). The resulting fingerprint is a string describing the frequency of every possible combination at predefined positions within the string. Several variants of pharmacophore fingerprints have been designed and are frequently used. Such a fingerprint can be used to analyze the similarity between molecules or among a library of molecules. Alternatively, a fingerprint model can be used to analyze the common elements of active ligands to identify the key contributing features to the biological function.

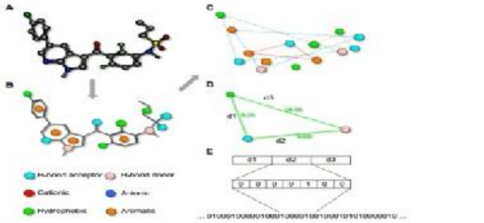


Figure 1 Pharmacophore fingerprint.
Notes: A pharmacophore fingerprint is the representation of a small-molecule ligand (A) annotated with molecular interaction features (B) into a string. Typically, every possible three- (or four-) point combination of molecular interaction features (C), with different distances between the features, calculated either through space or by the number of bond lengths (D), is calculated and the frequency of occurrence is stored in a string (E). Such strings are useful for the easy comparison of similarity between multiple molecules.

Figure 2

Sonophoresis-Activated Drug Delivery Systems

This type of activation-controlled drug delivery system utilizes ultrasonic energy to activate (or trigger) the delivery of drugs from a polymeric drug delivery device. The system can be fabricated from either a non-degradable polymer, such as ethylene–vinyl acetate copolymer, or a bioerodible polymer, such as poly [bis(p-carboxyphenoxy)alkane anhydride].The potential application of sonophoresis (or phonophoresis) to regulate the delivery of drugs was recently reviewed.

Pharmacophore model or query

A pharmacophore model consists of a few features organized in a specific 3D pattern.50 Each feature is typically represented as a sphere (although variants exist) with a radius determining the tolerance on the deviation from the exact position (Figure 2). The features can be labeled as a single feature or any logic combination consisting of "AND," "OR," and "NOT" to combine different interaction patterns within one label. Additional features can describe forbidden volume interactions (typically to represent the receptor boundary). Such pharmacophore features are typically used as queries to screen small molecule libraries of compounds.51 In these libraries all the compounds are present in their low-energy biorelevant conformations. Each of these con- formations is fitted to the pharmacophore query by aligning the pharmacophore features of the molecule and the query is composed. If a molecule can be fitted inside the spheres representing the query features it is considered a hit molecule. Often the pharmacophore query can be too complex to find hit molecules from a given library, and partial matching may be allowed. In such cases only certain features considered essential for activity are matched. Additional uses of such models are to align molecules or facilitate molecular dock- ing simulations.

Pharmacophore modeling in virtual screening:

Pharmacophore modeling is most often applied to virtual screening in order to identify molecules triggering the desired biological effect. For this purpose, researchers create a pharmacophore model (query) that most likely encodes the correct 3D organization of the required interaction pattern.

Depending on how much is known about the particular protein target, different options are available to construct such a query (Figure 3). In general, it is good practice to divide the ligand data into two sets, a training and an evaluation set to validate the generated pharmacophore query, when multiple active ligands (and inactive derivatives) are known.

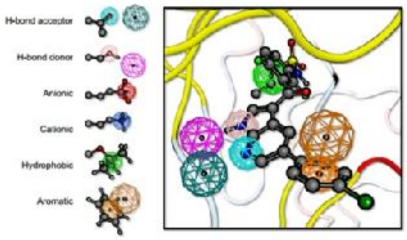


Figure 2 Pharmacophore query.
Notes: A pharmacophore query is comprised of different features. The features represent molecular recognition motifs such as hydrogen bond acceptors or donors, anionic, cationic, hydrophobic, and aromatic groups. The radius of the sphere determines the strictness of the geometric constraint. For features where the correct orientation of the interaction is important such as hydrogen bonds and the aromatic plane, a second feature can be used indicating the vector of the interaction (or the normal of the plane). A pharmacophore query can combine any of these features, with different radii and logic operations such as "AND," "OR," and "NOT." On the left a hypothetical pharmacophore query for BRAF kinase is given.

Figure 3

While in all these cases pharmacophore queries are considered positive filters to identify compounds, they may in fact also be used as negative filters to avoid side effects as well.

No protein structure and no ligand structure is known:

If the target structure and all its ligands are unknown, pharma- cophore modeling is impossible. The only option to employ the pharmacophore principle would be to design a diverse library employing a diversity metric based on pharmacophore fingerprints to ensure optimal diversity of the library, contain- ing a wide variety of molecules with different pharmacophore feature composition. Indeed, considering the large number of available and potential compounds, the trend is to design libraries very carefully in order to cover chemical space efficiently in any search process.

No protein structure, but active ligand structures are known:

The other scenario is that the structure of the receptor (and any complex with the ligand) is unknown. This is frequently the case in drug discovery. If only a single active molecule is known, then it is impossible to map the key contributing pharmacophore features onto the molecule, and the only option may be to use similarity searches (such as using phar- macophore fingerprints) to retrieve similar molecules.60 Once these have been tested, a set of multiple active and inactive compounds may be known and more advanced pharmacoph- ore modeling can be utilized.

When a set of active ligands of known structure, with similar or different scaffolds, is available, then it is possible to use ligand-based pharmacophore modeling. The elucidation of the putative pharmacophore involves two steps. First, the conformational space of the flexible molecules needs to be covered extensively since the bioactive conformations are unknown. Second, the molecules need to be aligned by com- mon pharmacophore features, which can be retained in a 3D model. Using inactive derivatives, the essence of the features as well as the permitted steric arrangement of the ligands can be mapped as well. The Catalyst-HypoGen algorithm in particular stands out from the variety of tools available for this purpose.62 This is a combination of QSAR and the pharmacophore method. It attempts to correlate structure and activity values (Ki or half maximal inhibitory concentra- tion [IC50]) by constructing a pharmacophore model. Thus, HypoGen not only identifies a query compound as "active" or "inactive" in the traditional function of a pharmacophore model, but also predicts activity value based on regression of the training dataset.

Protein and ligand structures are known:

In the third case, structural information is present for both ligands and the receptor protein. Usually a pharmacophore model represents the key features of a small molecule that allow it to bind to some receptor molecule, but this idea can be reversed and pharmacophore queries built from features of a protein active site. These features describe the principle interactions between the protein and its ligands, and can be mapped onto the bioactive conformation of the ligand. Ideally the structural model is derived from crystallographic or nuclear magnetic resonance data, but homology models or other structural data can be used as well. Although a struc- ture for one ligand may be enough, it is beneficial to have 3D information for multiple ligands to identify the common interactions. While this approach is compatible with the majority of pharmacophore modeling methods, LigandScout is notable as the first software package able to construct automatically a query from one or more Protein Data Bank (PDB) files based on protein–ligand interactions.64 Such structure-based pharmacophore queries have mul- tiple applications. They can be used for virtual screening, ligand binding pose prediction, and comparison of binding sites.

Only the protein structure is known:

In the last case, structural information for the protein receptor, but no active ligands, is known. In this case, a putative pharmacophore model can be constructed by analyzing the chemical properties of the binding site of interest. There are several different computational approaches that can directly convert 3D atomic structures of protein binding sites into queries. The interaction maps of the de novo drug design tool LUDI can be used to create a pharmacophore query.66 HS-Pharm is a knowledge-based method that uses machine-learning algorithms to prioritize the most interest- ing interacting atoms and to generate an interaction map within the binding site.67 Subsequently, the interaction map is converted into pharmacophore features. The GRID package is another approach to analyze the pocket in order to identify the key interactions.68 Using molecular interaction fields, the most favorable positions of atomic probes in the binding site can be identified and converted into pharmacophore features.69 Although many successes have been reported, the absence of any ligand structural information is a distinct disadvantage to drug design, since in the absence of a molecular scaffold it is hard to map the features in 3D space which can still be covered by atoms that are restrained by bond lengths and angles in the ligands.

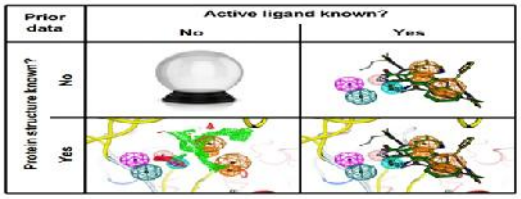


Figure 3 Four different situations for the pharmacophore search.
Notes: The figure shows the four different situations that may be encountered when starting a virtual screening. The situations include the absence of both the ligand and protein structure information, where except for divination, experimental screening is the only option. The second option is the presence of active ligands, but the protein structure is unknown, where pharmacophores can be used for ligand-based virtual screening. The best situation is when binding ligand and structural information is present. The most challenging option is when only a protein structure is available.

Figure 4

Pharmacophore methods in docking simulations:

As indicated in the previous section, pharmacophore models are very suitable as queries for virtual screening of databases. Nevertheless, one of the more common approaches in virtual screening is a so-called hierarchical approach in which differ- ent methods are combined consecutively. This is also known as the funnel principle, where at each consecutive step the compounds most unlikely to be active are removed, leav- ing the most promising compounds for virtual screening.72 Typically, every step of the hierarchical approach consists of a more complex, computationally demanding step than the previous one. As such, pharmacophore models are often utilized as a filter to identify compounds that fulfill simple geometric and chemical functionality requirements of the query, prior to more complicated and computationally demanding approaches such as molecular docking.

Molecular docking simulations are computational methods that aim to predict the binding mode of a compound for a given receptor as well as the quality of the interaction, often by attempting to predict the affinity (free energy of binding) using a scoring function.31 Often molecular docking simula- tions are used to screen large datasets of compounds for a given target, and compounds are ranked according to their predicted affinity. Due to the high number and diversity of the screening compounds, as well as the knowledge that most of the screened compounds are in fact probably inactive, the top scoring compounds are most likely inactive and better compounds are ranked below them. Although this ranking can still be better than random, typically only a few compounds are selected from those scoring best, and many of them often turn out to be inactive. Several options are available for combining docking- based virtual screening with pharmacophore-based virtual screening:

- ✓ The database of ligands can be pre-filtered using a pharmacophore query, prior to evaluation using docking simulations.
- ✓ The docking simulations can be post-filtered using a phar- macophore query to remove any compounds that fail to bind according to the pharmacophore query. The method can also discard compounds that would have scored well in a pure pharmacophore search, but that fail to bind according to some hypothesis taking more information into account, such as incompatibility of the overall ligand structure with the receptor site. In such a case, the ligands are evaluated in absolute conformation and should not be allowed to align with the pharmacophore features.
- ✓ Another alternative is to use the pharmacophore alignment to guide the placement during the docking simulations. The pharmacophore model can in this case be used for the placement of the ligand, similarly to the fitting of a molecule into the pharmacophore query; or to guide the placement by using a constraint while scoring the different docking poses. The pharmacophore query could originate from a user-defined query or an automatically generated receptor-based pharmacophore query.

Pharmacophore models are very useful for enriching the top scoring docking results with active compounds. This was demonstrated in the recent SAMPL4 virtual screening challenge where competitors were asked to rank a set of compounds for a given target, HIV-1 Integrase, without any Pharmacophore modeling in drug discovery knowledge of activity of the compounds in the library.The top results were obtained for the group using a hierarchical method consisting of pharmacophore pre-filtering as well as pharmacophore post-filtering of the docking results.

Future perspectives on pharmacophore modeling:

Pharmacophore modeling has been around since the beginning of CADD and has evolved from a basic concept into a well-established CADD method with applications including similarity metrics, virtual screening, ligand optimization, scaffold hopping, target identification, and so on. Given the simplicity and versatility of the pharmacophore concept, it can be anticipated that further developments will be made in the future for different applications.

Fragment-based drug design:

Over the last two decades, fragment-based drug design has become a well-established method for the rational development of novel drugs.101 Rather than screening drug-like molecules (with molecular weights of around 500 Da), smaller molecules with a molecular weight up to 350 Da (referred to as fragments) are being screened for affinity with a receptor using highly sensitive biophysical methods. Fragments showing some affinity for the target are grown into bigger and more potent compounds, and frag- ments binding to adjacent areas can be linked as well.

Since the diversity of small molecule fragments can easily be sampled with a few hundred compounds, in silico screening methods are highly suitable for fragment-based design. CADD methods such as docking and pharmacophore modeling have therefore also been used to identify fragment- like compounds in silico prior to testing in vitro; subsequent fragment recombination can be used for the de novo design of inhibitors..

Protein–protein interaction (PPI) inhibition:

Although once thought to be undruggable, “high-hanging fruits on the drug discovery tree,” PPIs have drawn a great deal of attention in recent years.The undruggable image has disappeared and an increasing number of small molecule inhibitors of PPIs (SMPPII) have been reported. Most of the early inhibitors originate from HTS.Structural analy- sis of proteins in PPI complexes and inhibitor complexes show that the interactions at the PPI interface are being mimicked by the ligand. SMPPII are found to copy the natural interaction not only in terms of shape and chemistry, but even at the electrostatic potential level.This mimicry suggests that the pharmacophore queries created from PPI complex structures can be used to identify SMPPII via virtual screening. Different methods can be employed to map the pharmacophore features onto the amino acids present at the PPI interface. Several SMPPII discoveries have been achieved, thanks to pharmacophore searches using manually created search features, or a consensus of interactions at the PPI interface, or using automated methods, or by identification of the key interactions using molecular interaction field analysis. PPIs are especially promising targets for controlling inappropriate signaling, as found in diseases such as cancer. The usefulness of pharmacophore modeling to create queries encoding the key interactions at the PPI interface will prob- ably strongly stimulate the discovery of novel SMPPII using pharmacophores, both as a stand-alone virtual screening tool and incorporated into pipelines with other methods.

A potential role in protein design?

Although pharmacophore modeling originated as a drug design concept and, as indicated earlier, is nowadays a key element of CADD, pharmacophore modeling shows promise in the currently burgeoning field of computational protein design. Rather than designing drugs for a given protein target, the aim in computational protein design is to derive an amino acid sequence that will fold into a given structure with a desired function. In many cases, this may involve protein– small molecule ligand interactions, and for these it can easily be imagined that pharmacophores may be used simply by reversing the process of small molecule drug design for a known protein structure.

First of all, suitable protein templates (enzymes or otherwise) should be identified for the protein redesign process. The ligand of interest could serve as a query to try to identify possible binding proteins, which can then later be redesigned to give optimum complementarity to the ligand. Second, during the virtual protein design process, often multiple rotamers of different amino acids are sampled to identify the most desirable ones.119 Similar to ligand fitting with a pharmacophore query, the protein side chains can be fitted to features describing the complementary interactions required at the protein–ligand interface.

3. Phramcophore Identification

Introduction:

The concept of a pharmacophore is widely used in modern drug design and it is generally defined as the 3D arrangement of certain features in the ligand that are responsible for its activity against a particular protein target. The importance of the pharmacophore stems from the fact that once it has been identified, it can be used to rationally design new ligands that contain it and thus have a greater chance of producing the desired pharmacological e ect. The pharmacophore can be relatively easily identified if the 3D structure structure is available for several ligand bound to the same binding site of the same the ligands and finding their largest common arrangement, referred to as the common pharmacophore (CP). The underlying assumption behind such an approach is that the structurally conserved characteristics of a set of active ligands are responsible for their biological activity against the specific protein target. However,the number of ligands for which the 3Dstructure has been determined representsavery small fraction of all the binding data currently available. In most cases such as inhigh-through put screening assays, the actual 3Dgeometry of the binding conformation is not known and only the activity is provided. In order to identify a common phar macophore in a set of ligands for which exact 3D structures are not known, one can usually enumerate all possible low-energy conformations of all ligands, identify a potential binding conformation for each ligand and then align those binding conformations to find the largest common arrangement of features. However, the identification of the binding conformation is not an easy task. It can be further complicated by the fact that different ligands may have different binding modes or bind to different sites of the target.

Methods

Definitions and Notations:

Each labeled and undirected graph G is represented as a tuple G = {V,E,L,L}, where V is a set of vertices or nodes, E ⊆V ×V is a set of undirected edges ofG, L is a set of disjoint vertex and edge labels, andL: V ∪E → L is a function that maps the vertices and edges to their corresponding labels. A graph is represented by its|V|×|V| adjacency matrix M in which each off-diagonal element Mi,j contains the label of the edge (vi,vj) and each diagonal element Mi,i contains the label of the vertex vi, Two graphs G1 = {V1,E1,L,L1} and G2 = {V2,E2,L,L2} are considered isomorphic if |V1| = |V2|, |E1| = |E2| and there exists a bijection f : V1 → V2 such that ∀v ∈ V1, L1(v) = L2(f(v)) and ∀(v,u) ∈ E1, L1((v,u)) = L2((f(v), f(u))) (i.e., there is a one-to-one correspondence of the vertices and edges between the two graphs). Graph G0 = {V0,E0} is called a subgraph of G = {V,E}, denoted by G0 ⊆G, if V0 ⊆ V and E0 ⊆ E. If there exists a subgraph g0 of a graph G that is isomorphic to a graph g, then g0 is called an embedding of g in graph G. IfagraphG contains atleast one embedding of a graph g, then g is said to be supported byG. A clique is a fully connected graph, i.e., for each pair of vertices in V there exist an edge in E. The size of a clique is defined by the number of vertices it contains, i.e., |V|. A clique with n vertices is called an n-clique. As a result, the number of edges in the n-clique is n×(n−1)/2. For example, Fig. 5 contains two graphs G1 and G2.

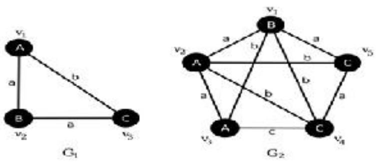


Figure 1: Example of graphs. G_1 is a clique, G_2 is not a clique.

Figure 5

Canonical Representation of Cliques:

Almost all graphs can be represented in more than one way depending on the order of vertices and edges in the representative string. This fact can significantly reduce the performance of most graph mining a lgorithms that will try to find the embeddings of the same graph multiple times due to multiple available representations. In order to avoid such redundancy, one needs to use a canonical graph labeling. Canonical label is the unique code of a given graph. The canonical code of a graph G (denoted by can (G)) should be the same regardless of its representation as long as the topological structure of the graph and its vertex and edge labels remain the same. The canonical code that we use is based on the minimum adjacency matrix code. This code is constructed by taking an adjacency matrix and rewriting it in one line by concatenating all its rows. The minimum adjacency matrix code isthe lexicographically minimum code among all possible adjacency matrix codes for a given graph. This canonical code has a prefix preservation property, i.e., for each graphG there exists a subgraphGs ⊆G such that the canonical code of the Gs is a prefix of the G's canonical code. In this study, we used a modified version of a minimum adjacency matrix code to represent the cliques to have all node labels in the code be in lexicographic order. The code is simply a string consisting of node labels followed by the edge labels to each preceding node in the clique in order. For example, the clique represented by graph G1 in Fig. 1 can have several codes: ABaCba≺ ACbBaa≺ BAaCab≺ BCaAab≺ CAbBaa≺ CBaAba. However, since the code ABaCba lexicographically precedes all other codes, it is the canonical code for that clique. Now, consider a clique formed by vertices v2, v3, and v4 and the connecting edges in graph G2 in Fig. 5. This clique can also be represented by several codes, two ofwhich, viz. AAaCbc and AAaCcb,have the same alphabetical order of the nodelabels. HoweversinceAAaCbc≺ AAaCcb, the AAaCbc is the canonical code for that clique.

Graph Representation of Conformations:

Each molecular conformation is represented as a graph, where pharmacophore points are the vertices and edges are the inter-point distances. From here on we will use the terms pharmacophore points and vertices as well as inter-point distances and edges interchangeably. For each graph G = {V,E,L,L}, there are two sets of labels in L = LV ∪ LE: one for vertices (LV) and another one for edges (LE). The labels of the vertices are used to capture the type of the pharmacophore points and they are user-defined. In the current study we used a total of six vertex labels corresponding to the following types of pharmacophore points: P (positive ionizable atom), N (negative ionizable atom),A(hydrogen-bondacceptor),D (hydrogen-bond donor), R (aromatic ring centroids), and H (hydrophobic). The labels of the edges are used to capture the distance between the pair of pharmacophore points associated with the vertices. In our graph model, the actual distances are discretized into a finite number of bins. Specifically, the distance range between [dmin,dmax] is discretized into l equal-size intervals with the corresponding labels 0,...,l−1. Based on the actual distance between a pair of pharmacophore points, each edge can be assigned up to two labels. The first label corresponds to the bin that its distance falls in. The second label is designed to maximize the identification of common pharmacophores due to the above distance binning. In particular, if the actual length of an edge is withinδ(whereδ≤0.5)of the b insize from a bin boundary, then the label corresponding to the adjacent left or right bin is also assigned to that edge. The parameters dmin,dmax, l, and δ are user-defined. Also, any distances that fall outside the [dmin,dmax] range are ignored and they are treated to represent vertices that are either too close to actually represent different pharmacophore points (e.g.,same atom was assigned more than one label) or they are too far away to be meaningful pharmacophore points. The multiple label assignment guarantees that all frequent cliques in which the variation of the distances be- tween any pairs of vertices does not exceed 2δ of the bin size will be discovered even if the corresponding distances in different molecules are assigned different labels. For example, if the bin size l = 1Å and δ = 0.25, then all of the distances differing by up to 2×δ×l = 0.5Å will be assigned at least one common label. Selecting a smaller valueof δ will decrease the run time but will only find very high-quality (low RMSD values) pharmacophores; selecting a larger value will find looser pharmacophores at the expense of more computational requirement. For example, selecting δ = 0.5 will always assign two labels to everyedgeandthealgorithmwillbeguaranteedtofindall cliques in which identical inter-point distances differ by no more than the bin size l.

Problem Definition:

Let D = {G1,...,Gn} be a set of sets of graphs, one for each of the active molecules{M1,...,Mn}. Each Gi is a set of graphs {Gi1,...,Gimi} for each of the mi conformations of the molecule Mi. For a given clique C, the supportofacliqueintheDisdefinedas sup(C) = |M|,where M⊆{M1,...,Mn}is a set of molecules that each have at least one conformer graph G that supports C. Thus, provided the above, the common pharmacophore identification problem is defined as follows: Given D and the minimum support fraction σ, find all cliques in D whose support is≥ σ•|D|. Setting σ to 100% restricts the search to only pharmacophores supported by all molecules. On the otherhand,allowing σ to be less than100% allows to find pharmacophores supported by only a portion of molecules which is required in cases when multiple binding sites, multiple binding modes or noisy data is a possibility.

Algorithms:

In this section we describe the two proposed algorithms for common pharmacophore identification using frequent clique mining in the graphs representing low-energy conformations of the active molecules, or ligands. The MCM algorithm is based on existing clique-mining methods and mines the conformer graphs using a depth-first approach and operating on edge-labeled graphs, and correctly determines the frequency of a common pharmacophore based on is embeddings in the pharmacophore graphs of the various ligands. The UCM algorithm improves the computational complexity of MCM by capitalizing on the fact that there is a high degree of structural similarity among the 3Dconformations of a molecule. Both algorithms produce identical results and differ only in the execution time.

MCM-Multiple Conformer Minor Logorithm

UCM-Unified Conformer Minor Logorithm

The MCM algorithm described in this section uses a depth-first approach to discover all frequent cliques. Duringeachstep, it generates a new candidaten+1-cliqueby growing the size of the current frequent n-clique via addition of a single new vertex and n new edges. All added edges are taken from the set of all frequent 2-cliques generated at the beginning of the algorithm. The clique is grown in such a way that the canonical code of the current clique is a prefix of the candidate clique. The latter ensures that the same clique is not enumerated multiple times. After the candidate generation step, the clique is enumerated (i.e.,the embedding so the clique are mined)and,if found frequent, reported and used for further growth. MCM algorithm uses a simple adjacency matrix M to store each conformer separately for each molecule. When a two-bin assignment is used, the two different labels (differing by 1) corresponding to the distance between the same two vertices are stored in Mi,j and Mj,i. If a single-bin assignment is selected, then all off-diagonal elements in the lower triangular matrix are assigned–1,i.e., ∀j < i : Mi,j = −1. Sample adjacency matrices for conformers depicted in Fig. 2 are presented in Fig. 3a (for visualization purposes, the numeric edge labels were replaced with alphabetical ones to avoid confusion with conformer IDs used by the unified conformer matrix de- scribed in the following section).

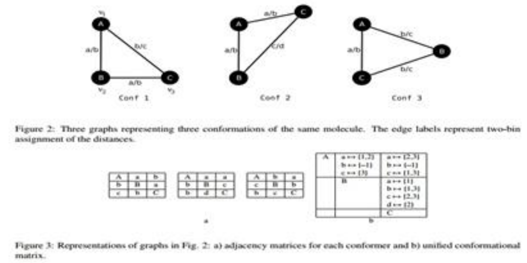


Figure 2: Three graphs representing three conformations of the same molecule. The edge labels represent two-bin assignment of the distances.

A	a	b
b	B	a
c	b	C

A	a	a
b	B	c
b	d	C

A	b	a
c	B	b
b	c	C

a

A	a ↔ {1,2}
	b ↔ {−1}
	c ↔ {3}
		a ↔ {2,3}
		b ↔ {−1}
		c ↔ {1,3}

	B	a ↔ {1}
		b ↔ {1,3}
		c ↔ {2,3}
		d ↔ {2}

		C

b

Figure 3: Representations of graphs in Fig. 2: a) adjacency matrices for each conformer and b) unified conformational matrix.

Figure 6

Pharmacophore model validation:

The main purpose of validating a quantitative model is to determinewhether our model is able to identify active structures and forecast their activity accurately. Therefore, two validation procedures were followed namely, test set prediction method and Cat-Scramble method.The compounds were used as test set to validate the pharmacophore model. The Cat-Scramble validation procedure is based on Fischer’s randomization test. The goal of this type of validation is to check whether there is a strong correlation between the chemical structures and the biological activity. This is done by randomizing the activity data associated with the training set compounds, generating pharmacophore hypotheses using the same features and parameters to develop the original pharmacophore hypothesis.

The statistical significance is calculated as following formula:

$$\text{Significance}=100(1-1+x/y)$$

Where x is the total number of hypotheses having a total cost lower than HypoX (original hypothesis), and y is the total number of HypoGen runs (initial + random runs). Thus 19 random spreadsheets (or 19 HypoGen runs) have to be generated for 95% confidence level. If the randomized data set results in the generation of a pharmacophore with similar or better cost values, RMSD, and correlation, then the original hypothesis is considered to have been generated by chance.

Database searching:

CATALYST generated best pharmacophore model comprising of four chemical features was used as a query for searching Maybridge chemical database consisting of 60,000 structurally diversifiedsmall molecules. Virtual screening of such databases can serve two main purposes: first, validating the quality of the generated pharmacophore models by selective detection of compounds with known inhibitory activity, and second, finding novel, potential leads suitable for further development. Best flexible search method was used for database searching to retrieve new lead molecules.

Homology modeling:

BLAST(blastp) was employed to search the relevant target or template proteins for building S.aureus MetRS protein structure. ClustalW multiple sequence alignment method was applied to compare the S. aureus MetRS sequence with other bacterial MetRS. The MODELLER module in INSIGHTII software was used to develop the homology model. Sequence alignments were achieved by Align2d method and the final 3D model was validated by PROCHECK software [26].

Molecular docking:

The program GOLD 3 (Genetic Optimisation for Ligand Docking) from Cambridge Crystallographic Data Center, UK[27] uses genetic algorithm for docking flexible ligands into protein binding sites to explore the full range of ligand conformational flexibility with partial flexibility of the protein. A pseudo atom was created at binding site region of modeled S. aureus MetRS whose coordinates were taken to define active site region with a active site radius of 8.0 A˚ . The annealing parameters of van der Waals and H-bond interactions were considered within 4.0 and 2.5 A ˚ , respectively.

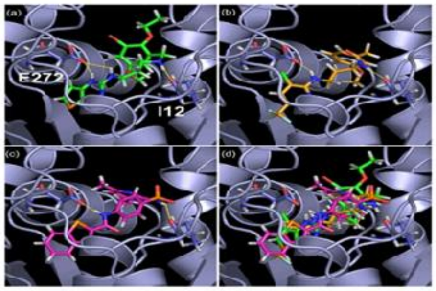


Fig. 9. The molecular docking results. The docked compound 1 of training set (a), Maybridge compound AW01179 (b), Maybridge compound BTB00521 (c), all three ligands at the binding pocket (d) are shown with the two catalytic residues (I12 and E272) of the modeled S. aureus MetRS structure.

Figure 7

4. Conclusion

The pharmacophore concept was first put forward as a useful picture of drug interactions almost a century ago, and with the rise in computational power over the last few decades, has become a well-established CADD method with numer- ous different applications in drug discovery. Depending on the prior knowledge of the system, pharmacophores can be used to identify derivatives of compounds, change the scaf- fold to new compounds with a similar target, virtual screen for novel inhibitors, profile compounds for ADME-tox, investigate possible off-targets, or just complement other molecular methods. While there are limitations to the pharmacophore concept, multiple remedies are available at any time to counter them. Given this versatility, it is expected that pharmacophore modeling will maintain a dominant role in CADD for the foreseeable future, and any medicinal chemist should be aware of its benefits and possibilities.

5. Acknowledgement

Xiaoyu Qing acknowledges Kuleuven IDO/11/008 for fund- ing, Xiao Yin Lee acknowledges Ajinomoto scholarship for ASEAN international students for funding, Joren De Raeymaeker acknowledges IWT 120050 project NEMOA for funding, and Arnout RD Voet acknowledges RIKEN for the FPR fellowship.

6. References

[1] Sastry SV, DeGennaro MD, Reddy IK, Khan MA. Drug Dev Ind Pharm 1997; 23 (2): 157– 165.

[2] Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. J Nat Prod. 2007;70(3):461–477.

[3] Lourenço AM, Ferreira LM, Branco PS. Molecules of natural origin, semi-synthesis and synthesis with anti-inflammatory and anticancer utilities. Curr Pharm Des. 2012;18(26):3979–4046.

[4] Wikberg JES, Spjuth O, Eklund M, Lapins M. Chemoinformatics Taking Biology into Account: Proteochemometrics. In: Guha R, Bender A, editors. Computational Approaches in Cheminformatics and Bioinformatics. Hoboken: John Wiley & Sons; 2011:57–92.

[5] PharmacophorePerception,Development,andUsein Drug Design; G¨uner, O. F., Ed.; International University Line: LaJolla, CA, 2000.4

[6] Pharmacophores and Pharmacophore Searches; Langer, T., Ho mann, R. D., Eds.; Wiley-VCH: Weinheim, 2006.

[7] Eglen, R. M.; Schneider, G.; B¨ohm, H.-J. High-Thoughput Screening and Virtual Screening: Entry Points to Drug Discovery. In Virtual Screening for Bioactive Molecules; B¨ohm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000; pp 1–14.

[8] VanDrie, J. H. Future Directions in Pharmacophore Discovery. In Pharmacophore Perception, Development, and Use in Drug Design; G¨uner, O. F., Ed.; International University Line: La Jolla, CA, 1999; pp 515–530.

[9] Martin, Y. C.; Bures, M. G.; Danaher, E. A.; De-Lazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. J. Comput.-Aided Mol. Des. 1993, 7, 83–102.

[10] Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. J. Chem. Inf. Comput. Sci. 1996, 36, 563–571.

[11] Jones,G.; Willett,P.;Glen,R.C.Ageneticalgorithm for flexible molecular overlay and pharmacophore

elucidation. J. Comput.-Aided Mol. Des. 1995, 9, 532–549.

[12] 11. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw D, Friesner RA. PHASE: A new engine for pharmacophore perception, DQSAR model development, and 3D database screening: 1. Methodology and preliminary results. J. Comput. Aided Mol. Des. 2006, 20, 647–671.

[13] Zhu, F.; Agrafiotis, D. K. Recursive Distance Parti- tioning Algorithm forv Common v Pharmacophore Identification. J. Chem. Inf. Model. 2007, 47, 1619–1625.

[14] Feng J, Sanil A, Young, SS, Pharmacophore Identification Using Gibbs Sampling. J. Chem. Inf. Model. 2006, 46, 1352–1359.

[15] Kuramochi, M.; Karypis, G. Frequent Subgraph Discovery; Proceedings of the IEEE International Conference on Data Mining (ICDM'01), San Jose, California,USA,November29-December2, 2001, IEEE Computer Society, 2001.