Chapter – IV

PHARMACOPHORE ANALYSIS
• Pharmacophore
• Pharmacophore modeling and analysis
• Cost analysis
IV.1 Pharmacophore

- “It is the ensemble of steric and electronic features”, that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response.\(^1\)
- The 3D arrangement of functional groups that enable a compound to exert a particular biological effect.
- The pharmacophore does not represent a real molecule or a real association of functional groups, but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds towards their target structure.\(^1\)

IV.1.1 Standard pharmacophore features

The definitions for the eleven standard pharmacophore features available in Discovery Studio are listed below:
HB Acceptor

This feature matches the following types of atoms or groups of atoms with surface accessibility:

- sp or sp$^2$ nitrogens that have a lone pair and charge less than or equal to zero
- sp$^3$ oxygens or sulfurs that have a lone pair and charge less than or equal to zero
- Non-basic amines that have a lone pair

HB Acceptor (lipid)

This feature matches the following types of atoms or groups of atoms:

- Nitrogens, oxygens, or sulfurs (except hypervalent) that have a lone pair and charge less than or equal to zero

HB Donor

This feature matches the following types of atoms or groups of atoms:

- Non-acidic hydroxyls
- Thiols
- Acetylenic hydrogens
- NHs (except tetrazoles and trifluoromethyl sulfonamide hydrogens)
Hydrophobic

This feature matches the following types of groups of atoms:

- A contiguous set of atoms that are not adjacent to any concentrations of charge (charged atoms or electronegative atoms), in a conformation such that the atoms have surface accessibility, including phenyl, cycloalkyl, isopropyl, and methyl

Hydrophobic (aliphatic)

This function is a proper subset of the hydrophobic function definition that includes only aliphatic atoms.

Hydrophobic (aromatic)

This function is a proper subset of the hydrophobic function definition that includes only aromatic atoms.

Negative charge

This feature matches the following types of atoms or groups of atoms:

- Negative charges not adjacent to a positive charge
**Negative ionizable**

This feature matches atoms or groups of atoms that are likely to be deprotonated at physiological pH, such as:

- Trifluoromethyl sulfonamide hydrogens
- Sulfonic acids (centroid of the three oxygens)
- Phosphonic acids (centroid of the three oxygens)
- Sulfinic, carboxylic, or phosphinic acids (centroid of the two oxygens)
- Tetrazoles
- Negative charges not adjacent to a positive charge

**Positive charge**

This feature matches the following types of atoms or groups of atoms:

- Positive charges not adjacent to a negative charge

**Positive ionizable**

This feature matches atoms or groups of atoms that are likely to be protonated at physiological pH, such as:

- Basic amines
- Basic secondary amidines (iminyl nitrogen)
- Basic primary amidines, except guanidines (centroid of the two nitrogens)
- Basic guanidines (centroid of the three nitrogens)
- Positive charges not adjacent to a negative charge
Ring aromatic

This feature matches aromatic rings with five or six member atoms.

IV.1.2 Criteria for a Satisfactory Pharmacophore Model

First, it has to highlight the functional groups involved in the interaction with the target, the nature of the non-covalent bonding and the different intercharge distances. This means that worthless images of ribbon and spaghetti models\textsuperscript{1-2}, without indication of the molecular features of the interacting partners, have to be avoided. This is true also for many unnecessary and opaque theoretical digressions. The model also has to show some predictive power and lead to the design of new, more potent compounds or, even better, of totally novel chemical structures, not evidently deriving from the translation of structural elements from one active series into the other. An interesting aspect of pharmacophore-based analogue design is referred to as scaffold hopping. It consists in the design of functional analogues by searching within large virtual compound libraries of isofunctional structures, but based on a different scaffold. The objective is to escape from a patented chemical class in identifying molecules in which the central scaffold is changed but the essential function-determining points are preserved and form the basis of a
relevant pharmacophore\textsuperscript{1,3}. The second criterion for a valid pharmacophore model is that it should discriminate stereo isomers. Stereo specificity is one of the principal attributes of pharmacological receptors and a perfect stereo chemical complementarity between the ligand and the binding-site protein is an essential criterion for high affinity and selectivity. A convincing example of enantiomeric discrimination was observed for GABA-A receptor antagonists\textsuperscript{1,4}. Sometimes a good pharmacophore model can explain apparently paradoxical observations, e.g. the unexpected affinity reversal found in R- and S-enantiomers of the sulpiride series on changing N-ethyl to N-benzyl derivatives.\textsuperscript{1,5} Finally, it has to account for the lack of activity of certain analogues of the active structures. The knowledge of structural or electronic parameters leading to poorly active or inactive compounds is a cost-lowering factor that allows the number of compounds to be synthesized to be reduced.\textsuperscript{1}

**IV.2. Pharmacophore Modeling and Analysis**

Rapidly identify promising new molecular entities with or without target-structured data with the help of Discovery Studio Pharmacophore Modeling and Analysis tools\textsuperscript{6}, which enable:

- Scaffold hopping
- Ligand- or structure-based pharmacophore model generation
Fragment-based and structure-based pharmacophore generation

Conformer generation and analysis

Lead identification and optimization

3D database building and searching

Geometric, descriptor-based querying

IV.2.1 Pharmacophore generation

Catalyst algorithms are used to support most of the pharmacophore related tasks in Discovery Studio. These tasks include automatic creation of pharmacophores, creation of ligand conformations, creation of compact, indexed ligand databases that can be searched by pharmacophores, and many others.7 Accelry's Catalyst can generate two types of chemical feature based models or hypothesis, depending on whether activity data or no activity10 CatHypo is a program for automatically generating pharmacophore models. There are two main algorithms used in CatHypo, HypoGen8 and HipHop9. The Figure IV.1 depicts the details of Pharmacophore model generation.

HypoGen attempts to derive SAR pharmacophore models from a set of molecules for which activity values on a given biological target have been measured.
HipHop is intended to derive common feature pharmacophore models using information from a set of active compounds.

In addition, both algorithms have been modified to incorporate excluded volumes into the pharmacophores. HypoGenRefine will attempt to add excluded volumes during the optimization phase of HypoGen in order to improve the overall correlation with activity and estimated activity. HipHopRefine is an algorithm that adds excluded volumes to any feature-based pharmacophore with location constraints. Typically this algorithm is used as a post-processor to the HipHop algorithm. HipHopRefine adds excluded volumes in regions that are occupied by inactive molecules only.

IV.2.2 HipHop – Qualitative pharmacophore model

HipHop identifies configurations or three-dimensional spatial arrangements of chemical features that are common to molecules in a training set. The configurations are identified by a pruned exhaustive search, starting with small sets of features and extending them until no larger common configuration is found. Training set members are evaluated on the basis of the types of chemical features they contain, along with the ability to adopt a confirmation that allows those features to be superimposed on a particular configurations. One can define how many molecules must map completely or partially to the pharmacophore. This option allows broader and more diverse
pharmacophores to be generated. The resultant pharmacophores are ranked as they are built. The ranking is a measure of how well the molecules map onto the proposed pharmacophores, as well as the rarity of the pharmacophore model. If a pharmacophore model is less likely to map to an inactive compound, it will be given a higher rank\textsuperscript{7,8,10,11}. Optionally excluded volumes to the pharmacophores based on information from a set of inactives.\textsuperscript{12}

**IV.2.2a Spreadsheet Set-up for HipHop**

1. *Molecules Hypothesis Generation Workbench imported into spreadsheet*

   Principal specifies the reference molecules reference configuration models are potential centers for hypothesis
   
   - If (0)- do not consider this molecules.
   - If (1)- consider configurations of this molecule.
   - If (2)- use this compound as a reference molecule used only for HipHop hypothesis generation.

2. MaxOmit Features specifies how many features for each compound may be omitted

   - If (0) - all features must map to generate hypothesis.
   - If (1) - all but one features must map to generated hypothesis.
   - If (2) - features need to map to generated hypothesis used only for HipHop hypothesis generation.
IV.2.2b HypoGen – Quantitative pharmacophore model

The HypoGen pharmacophore model identifies chemical functional features that are typical of active compounds, thus facilitating their differentiation from inactive compounds\(^{13,14}\) two training sets were selected by considering the structural diversity which represented covering a range of activities at least four orders of magnitude. HypoGen algorithm uses the data selected by you (that is the training set, conformational models, pharmacophore features, parameters and so on) to generate the top scoring predictive pharmacophores.

Three phases

Predictive pharmacophores are generated in three phases –

- constructive
- subtractive
- optimization phase

Constructive

The **constructive phase** identifies hypothesis that are common to the most active set of compounds. The most active set is determined by the following equation, “MA*UncMA-(A/UncA)>0.0”. Where MA is the activity of the most active compound. UncA is the uncertainty in the measured activity, and A is the activity of the compound. The most
active set of compounds is limited to a maximum of eight. Once the set is determined, HypoGen enumerates all possible pharmacophore features for each of the confirmations of the two most active compounds. Furthermore, the hypothesis must fit a minimum subset of the features of the remaining most active compounds in order to be considered. At the end of the constructive phase, a database of very large number of pharmacophore configurations is generated. The objective of the subtractive phase is to identify those pharmacophore configurations developed in the constructive phase that are also present in the least active set of molecules, and remove them. The first step is the identification of the least active set of molecules, and to remove them. This is accomplished by the use of equation

"\log(A) - \log(MA) > 3.5"

Where A is the activity of the current compound and MA is the activity of the most active compound. In simple terms, all compounds whose activity is 3.5 orders of magnitude less than that or the most active compound are considered to be in the set of least active molecules.

The value 3.5 is a user adjustable parameter, if needed (i.e., if the activity range of the dataset does not span more than 3.5 orders of magnitude).14-16
**Subtractive**

The subtractive process identifies the hypothesis that is common to the least active compounds. The least active set is determined by the following equation

“Log (Cmpdx)-log (Most Active Compound)>3.5”.

It enumerates all possible pharmacophore configurations. Then it checks for configurations shared with the most active compounds and eliminates if shared by more than half of the least active compounds leading to feasible pharmacophores.\textsuperscript{14-16}

**Optimization phase**

The optimization phase involves improvement of the hypothesis score. Small perturbations are applied to those pharmacophore configurations that survived the subtractive phase and are scored based on errors in activity estimates from regression and complexity of the hypothesis. The cost of a hypothesis is a quantitative extension of Occam’s razor. The total cost of each pharmacophore is computed by the sum of three costs: weight, error, and configuration. While the weight component increases with deviation of the feature weight from the ideal value of 2.0, the error component increases with RMS difference between the measured and estimated activities. The configuration cost is fixed and depends on the complexity of the pharmacophore upon completion of this phase.\textsuperscript{14-16}
IV.2.2c HypoGenRefine algorithm to add excluded volumes

The HypoGenRefine algorithm enhances the HypoGen algorithm to place excluded volumes in key locations in an attempt to model unfavorable steric interactions.

The HypoGenRefine algorithm makes two modifications to the HypoGen algorithm:

- The subtractive phase is eliminated. In HypoGen, the subtractive phase is used to identify and remove pharmacophores that exist in both the active and inactive molecules, since these pharmacophores cannot discriminate between active and inactive compounds. When attempting to incorporate steric effects, however, these are the potentially interesting pharmacophores. An inactive compound may contain the active pharmacophore, but be inactive due to unfavorable steric effects. Therefore, these pharmacophores remain as potential candidates in HypoGenRefine.

- New move types are added to the optimization phase. As a possible move type during the optimization, an excluded volume can be added or removed from the candidate Pharmacophore.¹⁴⁻¹⁶

When placing an excluded volume, HypoGenRefine inspects the volume space of the active and inactive molecules to locate reasonable positions for adding an excluded volume. Excluded volumes are placed using the following steps:

- Align the active molecules to the candidate pharmacophore.
- Align the inactive molecule that has the best fit to the candidate pharmacophore.
Find atoms in the inactive molecules that are at least a minimum distance from any atoms in the active molecules. These are the candidate points to add an excluded volume.

Select one of the points at random and place an excluded volume.

---

**Figure IV.1: Pharmacophore Model Generation**

**Mapping Analysis**

The HypoGen module makes the assumption that an active molecule should map more features than an inactive molecule. Therefore a molecule should be inactive because it either does not contain an important feature, or misses the feature as it cannot be orientated correctly in space. Based on this assumption the most active compounds should map all features of the hypothesis model. The
mapping information for all ten hypotheses for each compound in the training set is available in a log file, which also contains feature weights, tolerance values, type and location of each feature, as well as cost values.\textsuperscript{17-18} The mapping information shows the precise atom for each compound in the training set that maps onto the relevant pharmacophore feature. Where a feature fails to map on a compound an asterix appears. Where a hypothesis failed to match features for the most active compounds it was considered to be an inaccurate representation.

**IV.3. Cost Analysis during Hypothesis Generation**

During an automated hypothesis generation run, Catalyst considers and discards many thousands of models. It distinguishes between alternatives by applying a cost analysis. The overall assumption used is based on Occam’s razor principal, that between otherwise equivalent alternatives, the simplest model is best. Simplicity is defined using the minimum description length principle from information theory. That is, the simplest model is that which can be fully described using the tersest language. Catalyst uses bits for language, so the program assigns costs to hypotheses in terms of the number of bits required to describe them fully. The overall cost of a hypothesis is calculated by summing three cost factors, a weight cost, an error cost, and a configuration cost. These are qualitatively defined as:
**Weight cost**

A value that increases in a Gaussian form as the feature weight in a model deviates from an idealized value of 2.0. This cost factor is designed to favor hypotheses where the feature weights are close to 2.

**Error cost**

A value that increases as the RMS difference between estimated and measured activities for the training set molecules increases. This cost factor is designed to favor models where the correlation between estimated and measured activities is better.

**Configuration cost**

This is a fixed cost, which depends on the complexity of the hypothesis space being optimized. It is equal to the Entropy of the hypothesis space. Of these three, the Error cost factor has the major effect in establishing hypothesis cost.

During the beginning phase of an automated hypothesis generation, Catalyst calculates the cost of two theoretical hypotheses, one in which the Error cost is minimal (all compounds fall along a line of slope =1), and one where the Error cost is high (all compounds fall along a line of slope = 0). These models can be considered upper and lower bounds for the training set. The ideal hypothesis cost (Fixed cost) is reported in the
full file found in the hypothesis generation directory. This value tends to be 70-100 bits. The null hypothesis cost is reported in the .log file found in the same directory and is usually higher than the Fixed cost. The difference between these two costs is important. The greater the difference, the higher is the probability for finding useful model. In terms of hypothesis significance, what really matters is the magnitude of the difference between the cost of any returned hypothesis and the cost of the null hypothesis. In general, if this difference is greater than 60 bits, there is an excellent chance the model represents a true correlation. Since, most returned hypotheses will be higher in cost than the Fixed cost model, a difference between Fixed cost and null cost of 70 or more will be necessary in order to achieve the 60 bit difference. If a returned hypothesis has a cost that differs from the null hypothesis by 40-60 bits, there is a high probability it has a 75-90% chance of representing a true correlation in the data. As the difference becomes less than 40 bits, the likelihood of the hypothesis representing a true correlation in the data rapidly drops below 50%. Under these conditions, it may be difficult to find a model that can be shown to be predictive. In the extreme situation where the Fixed and null cost differential is small (< 20), there is little chance of succeeding and it is advisable to reconsider the training set before proceeding.
Another useful number is the Entropy of hypothesis space. This value is calculated early in the run and is in the .full file near the value for Fixed cost. The number is the exponent to the base 2 of the number of models Catalyst will attempt to optimize during the run. If this number is less than 18, a thorough analysis of all models will be carried out. If higher, Catalyst will truncate the list and some models will not be considered. The Entropy number also correlates with run time and resources used\textsuperscript{15}.

**IV.3.1 Hypothesis Validation**

One method of validating the hypothesis model generated is via an analysis of the cost of generating the pharmacophore. The greater the difference between the null cost and the total cost the more statistically valid the hypothesis, and thus, the greater the probability of this model being a true representation of the data. This analysis process was used for the all the pharmacophore models generated. These theoretical cost values are measured in units of bits. The null cost is the cost of generating a hypothesis where the error cost is high. The total cost is the actual cost of hypothesis generation, and the fixed cost is where the error cost is minimal or where the pharmacophore is perfect\textsuperscript{17,19-20}. 
If the difference between the total cost and the null hypothesis cost is more than 60 bits, there is greater than 90% probability that the model is a true representation of the data. If the difference is 40-60 bits, there is a 75-90% chance that it represents a true correlation of the data. When the difference becomes less than 40 bits, the probability of the hypothesis being a true representation rapidly falls below 50% and if the total-null cost difference is less than 20 bits there is little chance of it being accurate and the training set should be reconsidered.\textsuperscript{17-18}

\textbf{Figure IV.2} depicts the common features of the potential lead aurora B inhibitor (1 nM).
Figure IV.2: Common features of lead compound

Color code of standard pharmacophore features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrogen-bond acceptor</td>
<td>green</td>
</tr>
<tr>
<td>Hydrogen-bond donor</td>
<td>magenta</td>
</tr>
<tr>
<td>Hydrophobic</td>
<td>cyan</td>
</tr>
<tr>
<td>Negative charge</td>
<td>blue</td>
</tr>
<tr>
<td>Negative ionizable</td>
<td>blue</td>
</tr>
<tr>
<td>Positive charge</td>
<td>red</td>
</tr>
<tr>
<td>Positive ionizable</td>
<td>red</td>
</tr>
<tr>
<td>Aromatic ring</td>
<td>orange</td>
</tr>
</tbody>
</table>
References

1. Thierry Langer and Rémy D. Hoffmann, Pharmacophores and Pharmacophore Searches Edited, pp.29-30


6. Pharmacophore Modeling in Discovery Studio.


16. Automatic Pharmacophore Generation Using Catalyst Jon M. Sutter, Osman Güner, Rémy Hoffmann, Hong Li, Marvin Waldman

17. Synthesis, Molecular Modeling and Biological Activity of Methyl and Thiomethyl Substituted Pyrimidines as Corticotropin Releasing Hormone Type 1Antagonists. Adam McCluskey, Paul A Keller, Jody Morgan and James Garnera

