Chapter 7

Classification of Chemical Compound Pharmacophore Structures

Lynn Batten[1], C. Sean Bohun[2], Kell Cheng[3], Tom Doman[4], John Drew[5], Rod Edwards[2], Susan Kutay[3], Devon McCrea[2], Wendy Myrvold[2], Frank Ruskey[2], Joe Sawada[2], Pauline van den Driessche[2], Jana Vander Kloet[3], Kathryn L.B. Wood[2].

Report prepared by Andrej Bona[3] and Claude Laflamme[3].

7.1   Introduction

There is a staggeringly large number of chemical compounds that can be made; even when one considers only those hypothetical molecules that are "drug-like", there are still probably on the order of 10^{100} unique molecular structures. To date, only on the order of 10^7 compounds have been made and characterized. There is consequently no hope of producing the totality of known chemical compounds at any time in the future.

However, many of the compounds that have been synthesized, as well as those that could be imagined in the size 10^{100} set, are structurally close to each other and would be expected to interact with biological molecules such as enzymes (according to Fischer's "lock & key" concept) in a similar fashion. One promising notion of structural similarity is the concept of a "pharmacophore", where compounds matching the same pharmacophore often interact with biological molecules (enzymes, receptors, etc) in a similar way. For example, the pharmacophore fingerprint can provide valuable information in the search for novel active compounds, in the measure of diversity of a screening library, and of structure-activity relationships between these compounds.

We are also particularly interested in classifying low molecular weight chemical compounds for drug discovery projects that can inhibit the action of an enzyme by lodging themselves within its active site. Since the geometry of these compounds must match the configuration of the receiving protein, the pharmacophore information could be used to improve the success rate.

A pharmacophore is a structural abstraction of the interactions between various functional group types in a compound. They are described by a spatial representation of these groups as centres (or vertices) of geometrical polyhedra, together with pairwise distances between centres. The corresponding groups under consideration are typically *hydrogen bond acceptor* (HACC), *hydrogen bond donor* (HDON), *negative* (NEG) and *positive* (POS) charges, and *hydrophobic* (HYD) groups. Other groups can easily be incorporated into the analysis. The centre-to-centre distances have been estimated at between 2Å to 15Å,[7] but again, other distance ranges can be easily

---

[1] University of Manitoba
[2] University of Victoria
[3] University of Calgary
[4] Searle Corporation
[5] College of William and Mary
[6] University of Calgary
[7] 1Å $= 10^{-10}$m

adapted. Typically, the edge lengths are divided into distance intervals, where every pharmacophore with edge lengths in the same specified intervals yields similar chemical properties.

We provide an analysis that facilitates counting 3 and 4 centre pharmacophores, including a mathematical model for distance interval ratios, triangle and other inequality requirements for feasible triangles and tetrahedra, and symmetries. We also include some special cases as an indication of the sheer number of relevant pharmacophores even under very specific limiting circumstances.

Beside spatial symmetries and distance similarities for each edge of the polyhedra, there does not appear to be any other relevant structural similarity feature between two pharmacophores that can be used to reduce the classification of a typical compound.

7.2   Symmetries of the tetrahedron

To get very rough, but quite general estimates on the number of possible 3-centre and 4-centre pharmacophores, a well known mathematical counting principle called Burnside's lemma can be used. This formula counts the exact number of non-isomorphic triangles and tetrahedra with k possible edge types and c possible types of vertices (centres), taking into account equivalence classes of symmetries given by a specified group G acting on a triangle or tetrahedron, as given by:

$$T(k, c) = \frac{1}{|G|} \sum_{g \in G} \psi(g),$$

where $\psi(g)$ is the number of configurations fixed by g.

In the case of the 3-centre pharmacophore (represented by a triangle $\triangle$), symmetries correctly include mirror images. For the 4-centre case (represented by a tetrahedron $\boxtimes$), mirror images (a reflection through any one of the faces) are distinct in the chemical sense, and must be excluded from the computation.

The resulting estimate on the number of 3 and 4 centre pharmacophores is given by the following table:

symmetry type:	description	num ($\triangle$)	$\psi(g)$	num ($\boxtimes$)	$\psi(g)$
id	identity	1	$k^3 c^3$	1	$k^6 c^4$
(ab)(cd)	antipodal edges			3	$k^4 c^2$
(abc)	triangular face	2	$k^1 c^1$	8	$k^2 c^2$
(ab)	flip	3	$k^2 c^2$		
Total	$|G|$	6		12	

Table 7.1: Pharmacophores with 3 and 4 centres.

Note that for the tetrahedron each of the resulting permutations is even (the product of an even number of transpositions) and there are 12 of them. Thus the group is the alternating group, A_4. This fact is used later in the program as it computes canonical representatives of each tetrahedron.

For the triangle, the Burnside computation gives us

$$T(k, c)_\triangle = \frac{1}{6}(k^3 c^3 + 3k^2 c^2 + 2kc).$$

For the tetrahedron we get

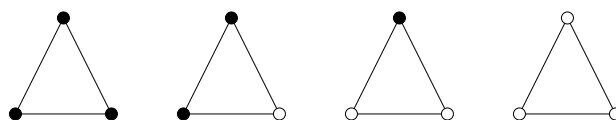$$T(k, c)_\boxtimes = \frac{1}{12}(k^6 c^4 + 3k^4 c^2 + 8k^2 c^2).$$

Here are listed different values of $T_\triangle$ and $T_\boxtimes$ for various values of k and c:

k	c	$T_\triangle(k,c)$	$T_\boxtimes(k,c)$
1	2	4	5
1	5	35	75
1	7	84	245
2	1	4	12
31	1	5456	74190161
31	5	632710	46229938075
31	7	1726669	177586039365

Table 7.2: Values of $T_\triangle$ and $T_\boxtimes$.

The actual configurations for $T(1,2)$ are shown below.

7.2.1 More than 4 centres

To create an $n + 1$-centre pharmacophore from an n-centre pharmacophore requires taking 3 additional edge distances and a particular face of the previously constructed polyhedron; this is sufficient to determine each new centre's location in space. The dominating term obtained from Burnside's lemma is therefore at most

$$\frac{1}{|G|}k^{3(n-2)}c^n,$$

where again n is the number of the centres, G is the group of automorphisms when these centres are placed in the "most symmetric way" on a sphere.[8] As an example, the dominating term in the number of 8-centre pharmacophores ($n = 8$, giving $|G| = 24$), with 30 possible edge types ($k = 30$), and 5 centre types ($c = 5$), is:

$$\frac{1}{24}30^{18}5^8 \simeq 6.3 \times 10^{30}.$$

7.2.2 Which distances give triangles and tetrahedra?

A more realistic count must consider the distances between centres, because an arbitrary set of distances is not necesarily achievable by triangles or tetrahedra. In the triangle case, a necessary and sufficient condition for numbers $a \leq b \leq c$ to be realized as side lengths is that $c \leq a + b$ (triangle inequality). For the tetrahedra, this condition holding on each triangular face is clearly necessary, but it is not sufficient. There are some necessary and sufficient conditions for distances to be achievable by a tetrahedron, but the following approach is more suitable for our purposes. Assume that two triangular faces have been determined — the triangles with sides a,b,c and b,d,e shown in Figure 7.1. Possible values that f can have must be determined.

The figure shows the two extreme cases for the length of f; in both cases the tetrahedron lies flat in the plane. Thinking of b as a hinge, all the possible tetrahedra with all distances except f fixed will be obtained. Noting that the angles are

$$\gamma = \cos^{-1}\left(\frac{a^2 + b^2 - c^2}{2ab}\right)$$

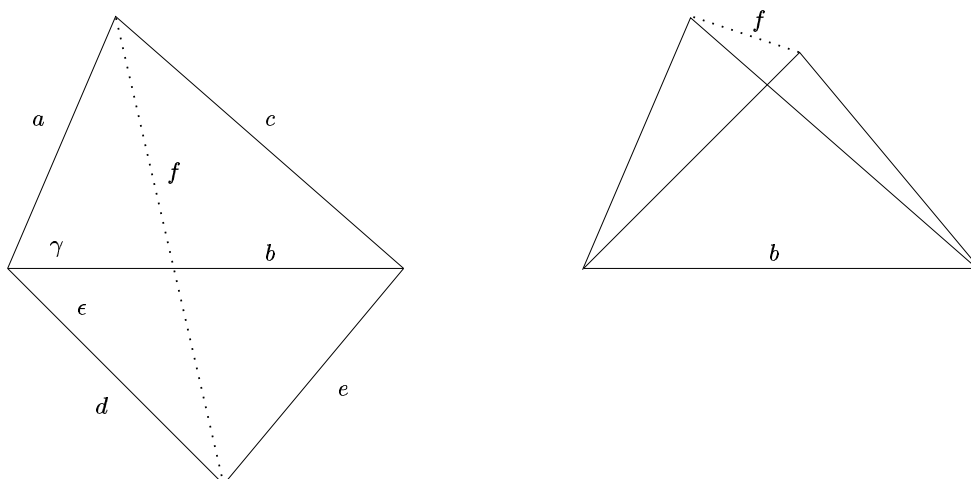[8]For chemical reasons we can assume that all centres lie on the convex hull of those centres (vertices).

Figure 7.1: Making a tetrahedron. Hinge is b. Extreme values for f shown.

and

$$\epsilon = \cos^{-1}\left(\frac{d^2 + b^2 - e^2}{2db}\right),$$

the possible values of f are bounded as follows

$$a^2 + d^2 - 2ad \cdot \cos(\gamma - \epsilon) \le f^2 \le a^2 + d^2 - 2ad \cdot \cos(\gamma + \epsilon). \tag{7.1}$$

Adding the triangle inequality constraints and (7.1) greatly reduces the number of possible pharmacophores. The exact magnitude of the reduction will depend on the number and lengths of the intervals on the edge lengths.

7.3   Chemistry

To further reduce our estimate of the number of pharmacophores, some possible chemical restrictions are considered in this section. A pharmacophore is a representation of the generalized molecular features that are considered to be responsible for a desired biological activity.[9] To improve the library of potential molecules of pharmacological activity, Searle is interested in developing a method that would improve the testing efficiency of drug-like molecules.

As stated in the introduction, the following five general categories of molecular features available for interaction with an enzyme are considered: hydrophobic [HYD], hydrogen bond donating [HDON], hydrogen bond accepting [HACC] and ionic sites which consist of positively [POS] and negatively [NEG] charged functional groups. Commercially, six to seven groups have been used to describe the active sites; however, we have constrained the number of pharmacophoric groups to five by grouping the aromatic and hydrophobic categories and by not considering acidic and basic functional groups as separate from the other features.

Lower bounds for the distances between the various possible combinations of these five types of sites need to be defined in terms of distances between the groups on a particular molecule based on bond lengths, bond angles and approximate degrees of rotation. When the active sites interact through space rather than the molecular structure of the compound, the sum of the van der Waals radii would be the limiting case in the calculation for the distance of closest approach. Here are listed some van der Waals radii for some important atoms: O: 1.4Å; N: 1.5Å; H: 1.2Å.

Pharmacophore space is defined by the relative distance between the centres of each site based on the binding opportunities of the enzyme active sites. The upper bound was defined by a reasonable distance (15Å) for the low molecular weight enzyme inhibitors. The basis for the distances are justified as follows:

---

[9]For more information please consult the web-site at `http://Intral.msi.com/weblab/medchem/mc_align.html`

- HDON or HACC site with another HDON or HACC site:

  Given the possibility of this molecular structure, two sites of the HDON or HACC type could be located on the hydrocarbon tails, which due to the flexibility of the tails, could approach distances as small as 2Å. Even a change in distance of 0.15Å could diminish or remove the beneficial interaction with the enzyme receptor because of the sensitivity of these interactions.
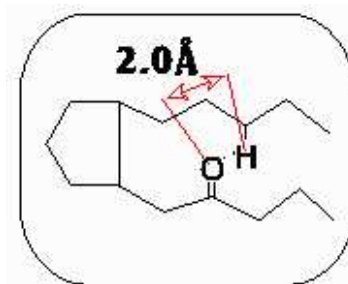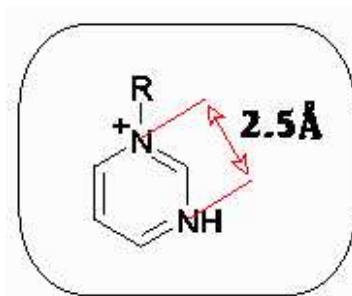

- An HDON - HACC site with a HYD site:

  The limiting distance between a hydrophobic centre and either an HDON or an HACC was determined by examination of a ketone/aldehyde site bonded to a hydrocarbon propyl chain. The centre of the HYD propyl group was based on the approximate centre of mass and the shortest distance between these sites was determined by the bond lengths and angles.
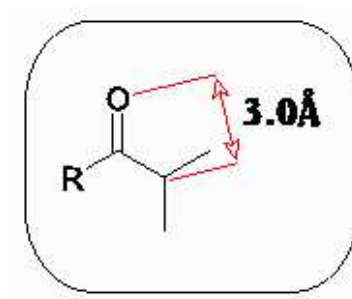
- An HDON - HACC site with a POS site:

  The closest approach between these two types of sites was found to be 2.5Å based on the bond angles and bond lengths of the ring illustrated. The figure shows that one nitrogen atom could behave as a positively charged site while the other could act as an HDON group.


- An HDON - HACC site with a NEG site:

  The closest approach between these two types of sites can be determined by the sum of the van der Waals radii. The example used to determined the distance of closest approach (3Å) was a the negatively charged oxygens of a phosphate group and a common HDON-HACC sites, such as hydroxide or ketone functional groups.

- An HYD side with another HYD site:

π

Given a sample molecule containing two hydrophobic groups, one a benzene ring and the other a propyl moiety, the approximate radius across the ring is 3Å and the approximate distance from the ring to the centre of propyl group is 1.5Å. The choice of a ring and a three carbon hydrocarbon chain as the model for a HYD-HYD interaction is justified since aromatic rings are common to drug molecules and the hydrocarbon chains with HYD functionality may contain as few as 3 carbons. The large margin for error of $\pm$ 1.0Å is acceptable since the hydrophobic interaction is diffuse and therefore not as directionally specific.
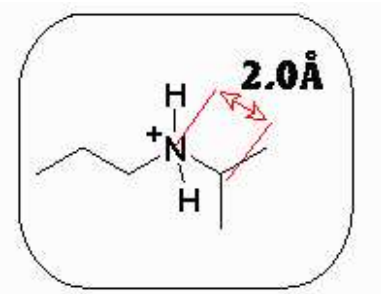
- An HYD site with a POS site:

  The shortest distance between a HYD and a POS site was determined under consideration of a POS amine group $(-NH_2^+-)$ bonded to a propyl HYD group. The distance was calculated based on the centre of the propyl group (as mentioned previously) and the position the nitrogen atom.
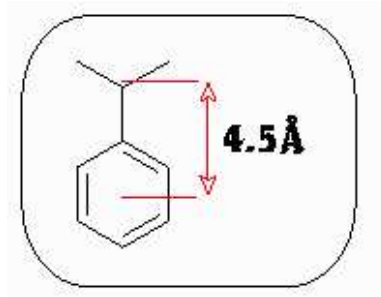
- An HYD site with a NEG site:

  The molecule illustrated here demonstrates a net negative charge at the COO^- group and the centre of the HYD group near the junction carbon of the propyl group. Thus, bond lengths (C-C) were used to approximate the shortest possible distance under these conditions.
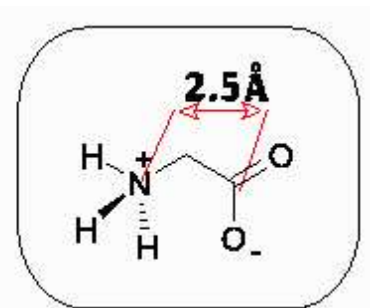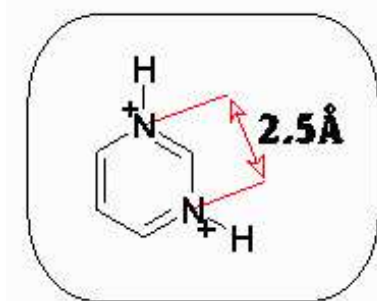
- A POS site with a NEG site:

  The shortest distance between these two centres was calculated based on the bond lengths of C-N and C-C and the angle of 109° between these bonds.
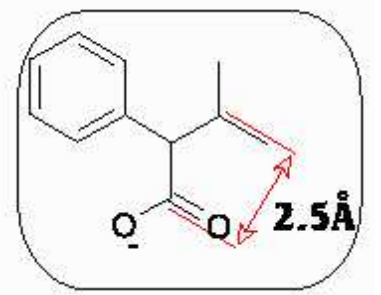
- A POS site with a POS site:

In the six member ring illustrated, both nitrogen atoms could accept a hydrogen atom, leaving them positively charged. This span was again calculated using C-N bond lengths and the angle of 120° between these bonds.

- A NEG site with a NEG site:

This common example of a functional group containing multiple negative charges has a larger minimal distance (3Å) between the various points of charge - the oxygen atoms - than the positive-positive charge distance (2.5Å). The sum of the van der Waals radii was utilized under these circumstances.
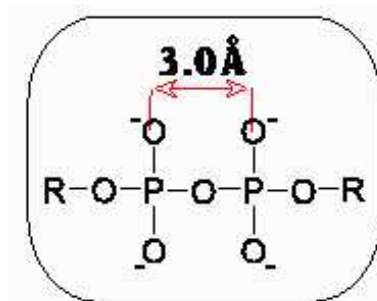
In the following table is given a summary of above results. Here is denoted the minimal distance between two active centres by d and minimal tolerance by δ.

	HDON/HACC	HYD	POS	NEG
HDON/HACC	0.15/2.0	0.25/3.0	0.25/2.5	0.25/3.0
HYD		1.0/4.5	0.25/2.0	0.25/2.5
POS			0.25/2.5	0.25/2.5
NEG				0.25/3.0

Table 7.3: Values of δ/d for various pairs of centres.

7.4 Bin Sizes

Two molecules define the same pharmacophore if they have the same configuration of centres and the set of distances between active parts of the molecules are relatively close. These differences may become larger with increasing distance between the centres in the sense that centres that are far apart should allow more freedom in the actual position of the centres involved. Ideally, we would like to index the different classes of pharmacophores. This suggests that the distances should scale as

$$\frac{dy}{y} = b\,dx$$

where b depends on the nature of the centres involved, y is the total distance between any two centres, dy is the maximum difference in distance that defines the same pharmacophore and dx is the difference between adjacent types. Solving for y we get

$$y(x) = a\,e^{bx} \tag{7.2}$$

π

for some value of a. This is a two parameter model and a and b can be determined by assigning to the first pharmacophore, $x = 1$, the minimal distance between the centers, d. As another constraint, the smallest allowable change in y at this minimal distance denoted as $\delta = y(2) - y(1)$ can be specified. The values of d and δ are empirical quantities which depend upon the chemical interaction involved. These empirical quantities are listed in Table 7.3.

Given a maximum distance between any two centres, D, the corresponding upper bound on the number of bins, N, can be computed by solving $y(N) = D$. This gives

$$N = \left\lfloor \frac{\ln(D/d)}{\ln(1 + \delta/d)} \right\rfloor + 1 \tag{7.3}$$

where $\lfloor \ \rfloor$ denotes the floor of the value. The last two columns of Table 7.4 list this value of N for each pair in Table 7.3.

Since the distance function (7.2) is exponential, a significant reduction in the value of N can be expected with only a slight change in the value of δ. However, just what is considered to be a slight change is open to interpretation and should depend strongly on the nature of the centres. Table 7.4 is included for demonstration of sensitivity of N with respect to δ/d.

δ/d	N	δ/d	N
0.1/2.0	42	0.15/2.0	29
0.2/3.0	26	0.25/3.0	21
0.2/2.5	24	0.25/2.5	20
0.2/2.0	22	0.25/2.0	18
1.0/4.5	7	1.0/4.5	7

Table 7.4: Values assume $D = 15$Å for all pairs.

7.5   Implementation

We have developed a program to estimate the number of three and four centre pharmacophores under user specified conditions. These include not only restrictions on the number of possible group labels, but also relative distance intervals and ranges for each of the pairs of centres. The program eliminates all duplicate pharmacophores due to possible symmetries in two or three dimensions. It also verifies the required conditions for feasible constructions of triangles and tetrahedra, taking into account margins of errors indicated in each of the distance bins.

Here are listed some of our results.
The final number of pharmacophores with:

- 3 centres, 5 labels (including exactly two HYD type centres) with a table of distances as in Table 7.3 is 3,704

- 3 centres, 5 labels (including exactly two HYD type centres) with a table of distances corresponding to first column of Table 7.4 is 5,775

- 3 centres, as previous, but any 5 centres is 295,535

- 4 centres, 5 labels (including exactly two HYD type centres) with a table of distances as in Table 7.3 is 27,953,097

- 4 centres with all properties as above, but instead of a min. increment for distance bins between HDON/HACC - HDON/HACC $\delta = 0.15$ we put $\delta = 0.12$, we get 29,941,835

- 4 centres, 5 labels (including exactly two HYD type centres) with a table of distances corresponding to first column of Table 7.4 is 88,178,007.

π

7.6 Conclusion

In this report is given an estimate on the size of the "pharmacophoric space" with some restrictions on types of possible pharmacophores. These restrictions were derived both from geometrical arguments and chemical properties of usual drug-like compounds. These estimates give a sense of the feasibility in building a screening library covering all 3 and 4 centre pharmacophores (to date there are only on the order of 10^7 compounds which are made and characterized), which could play an important role in drug searches especially if one can isolate extra restrictions in specific contexts.

A counting argument for five centre or larger pharmacophores would be very similar, but would involve a more delicate symmetry analysis of the pharmacophore. Rough estimates suggested by the calculations of Section 2.1 indicate that a library incorporating all 5 centre pharmacophores, for example, is not possible with current technologies. In these cases, new practical ideas to reduce the pharmacophore space are absolutely necessary.

π

Bibliography

[1] J. Greene, S. Kahn, H. Savoj, P. Sprague, S. Teig. Chemical Function Queries for 3D Database Search, *J. Chem. Inf. Comput. Sci.*, 1994, *34*, 1297-1308.

[2] D. Clark, G. Jones, P. Willett, P. Kenny, R. Glen. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching, *J. Chem. Inf. Comput. Sci.*, 1994, *34*, 197-206.

[3] S. Pickett, . Mason, I. McLay. Divesity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PQD), *J. Chem. Inf. Comput. Sci.*, 1996, *36*, 1214-1223.

[4] M. Ashton, M. Jaye, J. Mason. New perspectives in led generation II: Evaluating molecular diversity, *Drug Discovery Today*, 1996, *1*, No. 2, 71-78.

[5] M. McGregor, S. Muskal. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design, *J. Chem. Inf. Comput. Sci.*, 1998.