University of Antwerp

Chemo-informatics and computational drug design

Prof. Dr. Hans De Winter

University of Antwerp

Campus Drie Eiken, Building A

Universiteitsplein 1, 2610 Wilrijk, Belgium

Gefinancierd door de Europese Unie
NextGenerationEU

Chapter 8. Virtual screening

1. Pharmacophore searching

1.1. What is a pharmacophore?

A pharmacophore is an abstract description of molecular features that are necessary for molecular recognition of a ligand by a biological macromolecule. IUPAC defines a pharmacophore to be 'an 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'. A pharmacophore is defined as an ensemble of these interactions, or more specific the corresponding chemical features and their relative positions and orientations. It can be seen as a powerful abstraction or representation of small molecule binding to proteins. Essential interactions, corresponding to chemical features, are hydrogen bonding, charge transfer, steric and electrostatic characteristics, and lipophilic interactions. The strength of feature-based pharmacophore models lies in the adequate definition of the pharmacophore points

1.2. Pharmacophore features

Typical pharmacophore features include hydrophobic centers, aromatic rings, hydrogen bond acceptors and donors, positive charges and negative charges. Combinations of these are also possible (Figure 76).

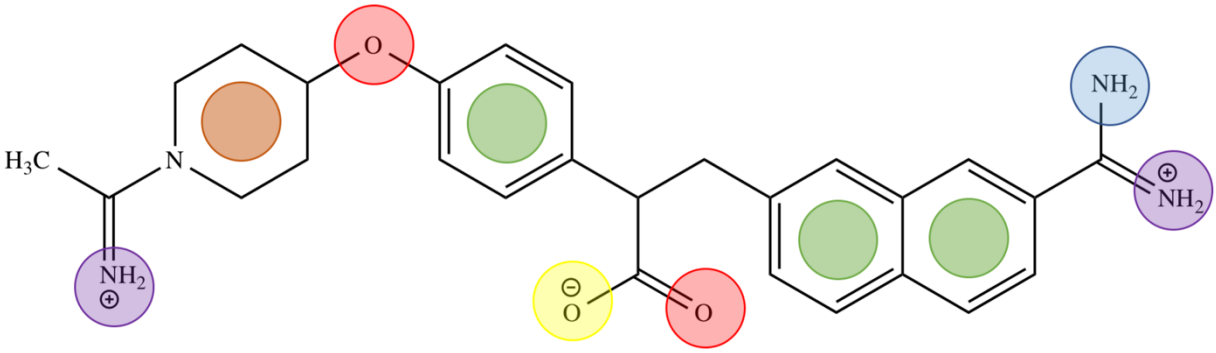

Figure 76. Schematic representation of a pharmacophore model of a ligand. The different pharmacophore types or features are indicated by different colors: purple, positive charge center and hydrogen bond donor; blue, hydrogen bond donor; brown, lipophilic; yellow, negative charge center and hydrogen bond acceptor; red, hydrogen bond acceptor; green, aromatic and lipophilic.

The program Pharao is a widely used pharmacophore searching program [10]. Pharao uses the following pharmacophore point definitions:

- *Aromatic rings*: the identification of aromatic ring pharmacophore points includes ring detection and aromaticity detection. A ring is labeled as aromatic if it is planar, has no exocyclic double bonds and satisfies Hückel's 4n + 2 rule;
- *Hydrogen bond donors*: hydrogen bond donor pharmacophore points correspond to atoms fulfilling the following conditions:
    - atom is nitrogen or oxygen;
    - formal charge of atom is not negative;
    - atom has at least one covalently attached hydrogen atom.
- *Hydrogen bond acceptors*: for the generation of hydrogen bond acceptor points, the following criteria need to be met:
    - atom is nitrogen or oxygen;

- formal charge of atom is not positive;
- atom has at least one available 'lone pair';
- atom is 'accessible'.

- *Charge centers*: atoms with a positive charge will correspond to a positive charge pharmacophore point, while atoms with a negative charge will correspond to a negative charge pharmacophore point.
- *Lipophilic spots*: to generate lipophilic pharmacophore points, a three-step procedure based on the method of Greene et al. [13] is used. First, every atom is assigned a 'lipophilic contribution'. This value is the product of a topology-dependent term *t* and the accessible surface fraction *s*. Term *t* is obtained using some simple heuristic rules that are based on the atom type, and fraction *s* is calculated with a solvent-accessible surface algorithm. Second, when a lipophilic contribution has been assigned to every atom, the next step is to group atoms into lipophilic regions or spots. Grouping atoms into spots is a simple procedure: (1) atoms in a ring of size 7 or less form a group; (2) atoms with three or more bonds, together with their neighbors and not bonded to any other non-hydrogen atom, form a group; (3) the remaining atoms are divided in chains on the basis of their connectivity, and each chain is defined as another group. Rings larger than seven atoms also count as chains. In the third step the lipophilic contribution for every spot is calculated as the summation of the contributions of the individual atoms belonging to that group or spot. If this value exceeds a predefined threshold, the spot corresponds to a lipophilic pharmacophore point for which the center coincides with the geometric center of this spot.

Combination of these six features results in a set of eight different pharmacophore types:

- AROM: aromatic ring
- HDON: hydrogen bond donor
- HACC: hydrogen bond acceptor
- LIPO: lipophilic region
- POSC: positive charge center
- NEGC: negative charge center
- HYBH: HDON + HACC
- HYBL: AROM + LIPO

Each pharmacophore point is represented by its type (eight possibilities as listed above) and a volume, representing the size of the pharmacophore point (Figure 77). The volume is modeled as a Gaussian 3D volume, calculated as:

$$V = \int e^{\left(-\frac{|\vec{m}-\vec{r}|^2}{\sigma}\right)} dr$$

with $\vec{m}$ as the center of the pharmacophore, and σ its spread (related to the volume).

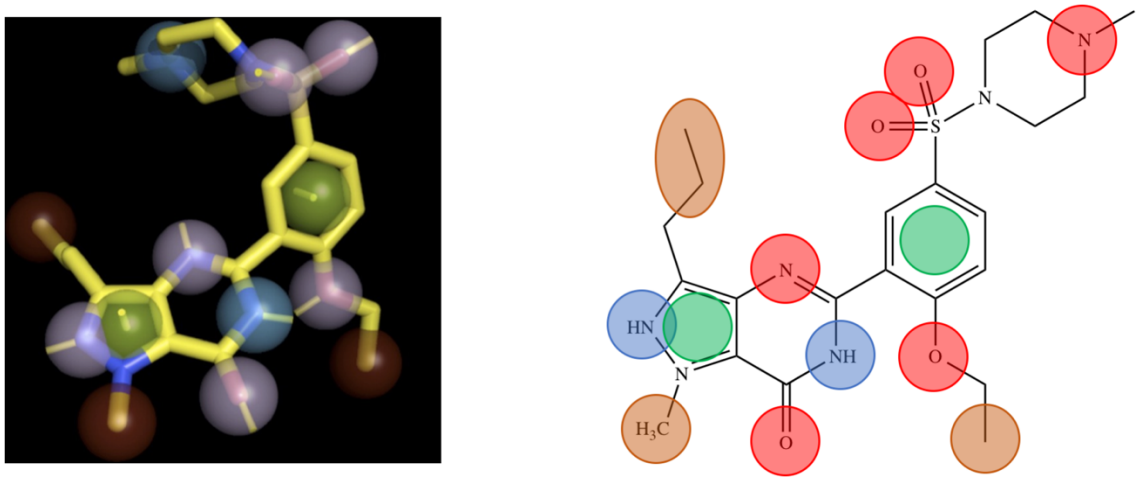

Figure 77. Example of a pharmacophore. The pharmacophore contains 14 pharmacophore points: two AROM points, three LIPO points, two HDON points and six HACC points.

1.3. Pharmacophore alignment

The quantification of the similarity between two pharmacophores can be computed from the overlap volume of the Gaussian volumes of the respective pharmacophores. The goal is to find the subset of matching functional groups in each pharmacophore that gives the largest overlap.

The procedure to compute the volume overlap between two pharmacophores is done in two steps. In the first step, a list of all feasible combinations of overlapping pharmacophore points is generated. In the second step, the corresponding features are aligned with each other using an optimization algorithm. The combination of features that gives the maximal volume overlap is retained to give the resulting score.

Step 1. Feature mapping

To compute the overlap between two pharmacophores, the first step is to define which points from the first pharmacophore can be mapped onto points from the second pharmacophore. A mapping of two pharmacophores consists of a list of points both pharmacophores where corresponding points have a compatible functional group and the internal distance between points is within a given range requirement. This range, as defined by the parameter ε, controls the feasibility of a combination of pharmacophore points.

The procedure starts by generating a list of all feasible pairs of features. First, two points from the first pharmacophore are selected (*a* and *b*) and the distance between them is calculated (d_{ab}). Next, two points with matching features are selected from the second pharmacophore (*c* and *d*) and the distance between these two points is also calculated (d_{cd}). The difference between the two distances is then compared to the sum of the sigmas of the four pharmacophore points. If

$$\varepsilon < \frac{|d_{ab} - d_{cd}|}{\sigma_a + \sigma_b + \sigma_c + \sigma_d}$$

then the combination of the two pairs is said to be feasible. This is also illustrated in Fig. 2. When ε is set to 1.0, this relates to the hard sphere atom model where the spheres are only touching each other and do not overlap. Smaller values of ε indicate that more overlap is required and becomes as such a more stringent selection criterion. Normally, ε is set to 0.5.

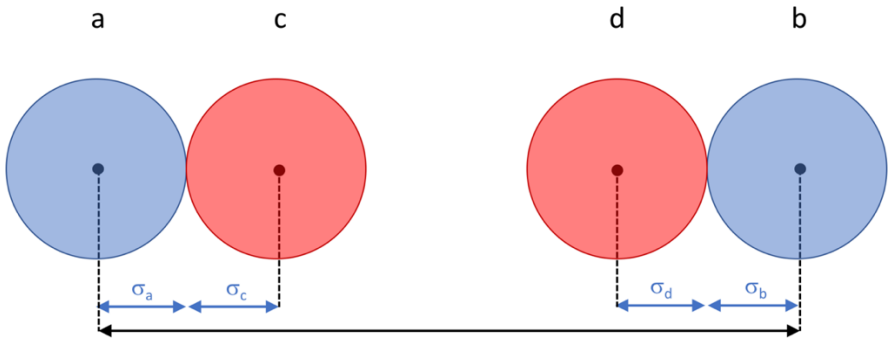

Figure 78. Illustration of the ε parameter. In this example, the difference between d_{ab} (black line) and d_{cd} (red line) is equal to the sum of the four sigmas (blue lines) of the pharmacophore points. When ε is smaller than 1.0, this implies that $|d_{ab} - d_{cd}|$ should be smaller than the sum of the four sigmas and thus the pharmacophore points should overlap.

Once the list of feasible pairs is constructed, they can be combined into larger feasible combinations. A combination of *n* pairs can be extended with any other pair if that pair is feasible and compatible with all the *n* pairs of the combination. This process is combinatorial in nature and the number of possible combinations grows very fast with the number of pharmacophore points. The ε parameter leads to a reduction of the number of feasible combinations.

Step 2. Alignment phase

Starting from the set of feasible combinations, the combination that gives the largest volume overlap is searched for. For every combination, the procedure starts by orienting the first and second pharmacophore subsets such that their geometric centre and their principal axes of inertia coincide. Next, by applying a constrained gradient-ascent to the rigid-body rotation of the second pharmacophore, the maximal volume overlap is determined between the two subsets. The rotational part is implemented using quaternion algebra. The use of the Gaussian representation of pharmacophore points offers an elegant way to compute the gradient and Hessian of the volume overlap. The result of the optimization procedure is the rotational angle and axis that gives the optimal overlap, and an alignment score which quantifies this overlap.

The complete alignment procedure starts from the subset with the largest number of matching points and computes the optimal volume overlap of this combination. Next, the smaller combinations are processed until the highest volume overlap so far is higher than the maximum volume overlap any smaller combination hypothetically can achieve. The rationale is that the volume overlap has an upper boundary that depends on the number of features to align. If the current best overlap is larger than this upper bound then there is no need to compute the alignment of smaller subsets since the score will never be larger than the current best.

Calculating the similarity between a pair of pharmacophores, a number of different measures can be used depending on the desired outcome. The most important are:

$$TANIMOTO = \frac{V_{overlap}}{V_1 + V_2 - V_{overlap}}$$

$$TVERSKY = \frac{V_{overlap}}{V_1}$$

with $V_{overlap}$ being the volume overlap of the matching subset of pharmacophores points, V_1 the volume of the first pharmacophore, and V_2 the volume of the second pharmacophore. The *TANIMOTO* measure is well known from bit vector comparison and is the default measure to score similarity between pharmacophores. The *TVERSKY* measure is to identify compounds having a pharmacophore that is a superset of the first pharmacophore (V_1). All two metrics are returning a score between 0 and 1.

2. Docking

Docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using a scoring function. Molecular docking is one of the most frequently used methods in structure-based drug design, due to its ability to predict the binding-conformation of small molecule ligands to the appropriate target binding site. Characterisation of the binding behaviour plays an important role in rational design of drugs as well as to elucidate fundamental biochemical processes.

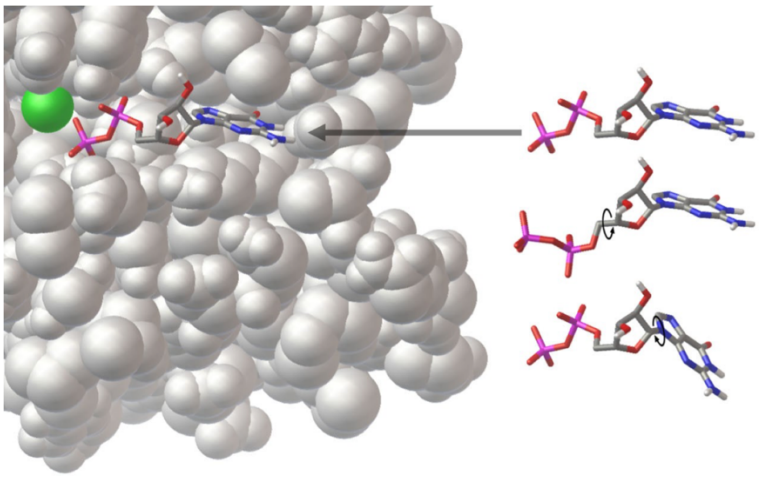

Figure 79. One can think of molecular docking as a problem of 'lock-and-key', in which one wants to find the correct relative orientation of the 'key' which will open up the 'lock'. Here, the protein can be thought of as the 'lock' and the ligand can be thought of as a 'key'. Molecular docking may be defined as an optimization problem, which would describe the 'best-fit' orientation of a ligand that binds to a particular protein of interest.

Molecular docking research focusses on computationally simulating the molecular recognition process. It aims to achieve an optimized conformation for both the protein and ligand and relative orientation between protein and ligand such that the free energy of the overall system is minimized (Figure 79).

To perform a docking screen, the first requirement is a structure of the protein of interest. Usually the structure has been determined using a biophysical technique such as x-ray crystallography or NMR spectroscopy, but can also derive from homology modeling construction. This protein structure and a database of potential ligands serve as inputs to a docking program. The success of a docking program depends on two components: the search algorithm and the scoring function (Figure 80).

Figure 80. Docking is the repeated process of searching and scoring, until certain predefined criteria are met.

2.1. Search algorithms

The search space consists of all possible orientations and conformations of the protein paired with the ligand. However, in practice with current computational resources, it is impossible to exhaustively explore the search space - this would involve enumerating all possible distortions of each molecule (molecules are dynamic and exist in an ensemble of conformational states) and all possible rotational and translational orientations of the ligand

relative to the protein at a given level of granularity. Most docking programs in use account for the whole conformational space of the ligand (flexible ligand), and several attempt to model a flexible protein receptor. Each 'snapshot' of the pair is referred to as a pose.

A variety of conformational search strategies have been applied to the ligand and to the receptor. These include:

- Shape-complementarity methods
- Molecular dynamics simulations
- Genetic algorithms

Shape-complementarity methods

As being the most common technique used in many docking programs, shape-complementarity methods focus on the match between the receptor and the ligand in order to find an optimal pose. Programs include:

- DOCK (http://dock.compbio.ucsf.edu)
- FRED (https://docs.eyesopen.com/oedocking/fred.html)
- GLIDE (https://www.schrodinger.com/glide), and
- SURFLEX (https://omictools.com/surflex-dock-tool).

Most methods describe the molecules in terms of a finite number of descriptors that include structural complementarity and binding complementarity. Structural complementarity is mostly a geometric description of the molecules, including solvent-accessible surface area, overall shape and geometric constraints between atoms in the protein and ligand. Binding complementarity considers features like hydrogen bonding interactions, hydrophobic contacts and van der Waals interactions to describe how well a particular ligand will bind to the protein. Both kinds of descriptors are conveniently represented in the form of structural templates which are then used to quickly match potential compounds that will bind well at the active site of the protein.

As one of the most established docking programs in the last decades, DOCK will be used as an example to explain the principles behind shape-complementarity methods. The DOCK procedure consists of three steps [11]:

1. Representation of the receptor and ligand structures;
2. Matching of the receptor and ligand representations;
3. Optimization of the ligand within the binding site.

The **representation step** consists of generating a set of spheres that fill all pockets and grooves on the surface of the receptor molecule. These spheres are collected into a number of presumptive 'binding' sites, and each site is then examined for geometric matching with the ligand. The ligand molecule is also represented by a set of spheres that approximately fill the space occupied by the ligand.

The molecular surfaces of receptor and ligand are required for this step. This surface is a collection of points and the vectors normal to the surface at each point (Figure 81).

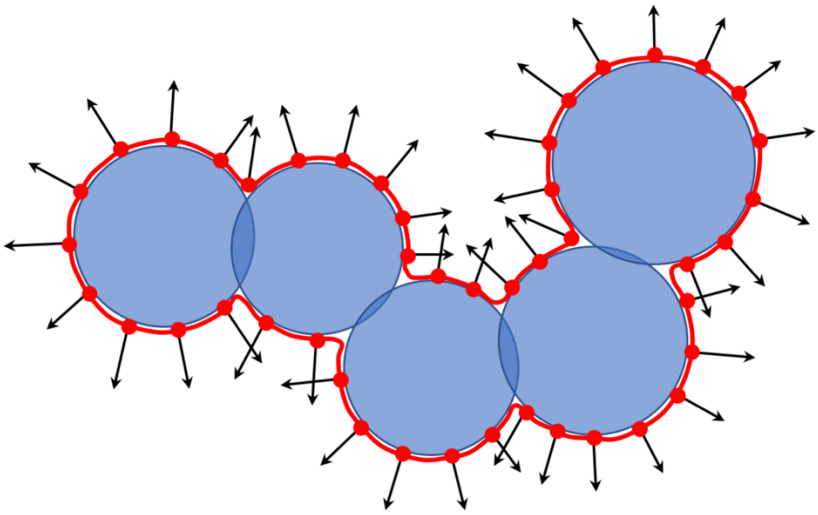

Figure 81. The blue spheres describe a receptor composed of five atoms, the red line the surface of the receptor, and the red dots are a set of points equally distributed on the surface. The black arrows are the normals of the surface at each point.

To reduce the number of points on the surface area, a more compact representation is then generated by constructing a set of spheres with the following properties:

a) Each sphere touches the molecular surface at two points (*i*, *j*) and has its centre on the surface normal from point *i* (Figure 82);

b) Each receptor sphere lies on the outside of the surface and each ligand sphere lies on the inside of the surface.

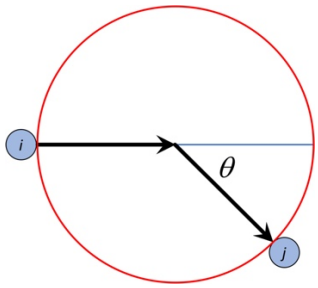

Figure 82. Sphere generated tangent to the surface points i and j with its center on the surface normal at point i.

It can be shown that, for a surface composed of *n* points, *n*-1 spheres can be constructed at each of the surface points. Therefore, the spheres that are generated in this way are reduced in two ways. First, of each of the *n*-1 spheres at each surface point, only the sphere with the smallest radius is kept. Second, for each smallest sphere, the angle formed by the two surface points, *i* and *j*, and the sphere centre, is calculated (Figure 82), and only those spheres are retained for which the angles are less that 90°.

The **matching rule** is based on the comparison of the internal distances in both receptor and ligand. A ligand sphere can be paired with a receptor sphere if the internal distances of all ligand spheres can be matched with all the internal distances of the receptor set, within some error limit on each distance. However, the complexity of the matching problem is that not all combinations can be tried, and for this reason a more pragmatic approach has been taken. First, each ligand sphere *i* is paired with each receptor sphere *k*, and all distances between ligand sphere *i* and the remaining ligand spheres *j* (d_{ij}) is calculated, as well as all distances between receptor sphere *k* to all remaining receptor spheres *g* (d_{kg}). A second pair of spheres ($j = g$) is assigned so that a maximum number of spheres obey the condition:

$$\left|d_{ij} - d_{kg}\right| < 1$$

in which the number '1' is a distance cut-off that has been chosen experimentally.

Once the best second pair of ligand-receptor spheres has been identified, a third pair of spheres is selected subject to the additional constraint that the distances from the new spheres to the previously assigned pairs must also obey the error check. This procedure continues until no further pairs can be assigned.

In the third phase, the **optimization stage**, all suggested pairings are explored. It carries out the rotation of the ligand spheres onto the corresponding receptor spheres using a least-squares minimization algorithm. The generated rotation/translation matrix is then applied to all ligand atoms and the resulting conformations are subjected to a user-defined scoring function for ranking and classification (see further).

Molecular dynamics simulations

In this approach, proteins are typically held rigid, and the ligand is allowed to freely explore their conformational space. The generated conformations are then docked successively into the protein, and an MD simulation consisting of a simulated annealing protocol is performed. This is usually supplemented with short MD energy minimization steps, and the energies determined from the MD runs are used for ranking the overall scoring. Although this is a computer-expensive method (involving potentially hundreds of MD runs), it has some advantages as no specialized energy/scoring functions are required. MD forcefields can typically be used to find poses that are reasonable and can be compared with experimental structures.

Because in this approach the protein is kept rigid, one might miss important protein conformations that are involved in ligand binding. One approach that aims to address this issue is ensemble-based docking. With this technique, ligands are docked to an ensemble of rigid protein conformations. Molecular dynamics simulations are used to generate the ensemble of protein conformations for the subsequent docking.

Genetic algorithms

Two of the most used docking programs belong to this class of docking approaches:

- GOLD (https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/)
- AutoDock (http://autodock.scripps.edu)

Genetic algorithms allow the exploration of a large conformational space - which is basically spanned by the protein and ligand jointly in this case - by representing each spatial arrangement of the pair as a 'gene' with a particular energy. The entire genome thus represents the complete energy landscape which is to be explored. The simulation of the evolution of the genome is carried out by cross-over techniques similar to biological evolution, where random pairs of individuals (conformations) are 'mated' with the possibility for a random mutation in the offspring. These methods have proven very useful in sampling the vast state-space while maintaining closeness to the actual process involved.

Although genetic algorithms are quite successful in sampling the large conformational space, many docking programs require the protein to remain fixed, while allowing only the ligand to flex and adjust to the active site of the protein. Genetic algorithms also require multiple runs to obtain reliable answers regarding ligands that may bind to the protein. The time it takes to typically run a genetic algorithm in order to allow a proper pose may be longer, hence these methods may not be as efficient as shape complementarity-based approaches in screening large databases of compounds. Recent improvements in using grid-based evaluation of energies, limiting the exploration of the conformational changes at only local areas (active sites) of interest, and improved tabling methods have significantly enhanced the performance of genetic algorithms and made them suitable for virtual screening applications.

2.2. Scoring functions

The search algorithms of the docking programs generate a large number of potential ligand poses, of which some can be immediately rejected due to clashes with the protein. The remainder are evaluated using some scoring function, which takes a pose as input and returns a number indicating the likelihood that the pose represents a favourable binding interaction and ranks one ligand relative to another. Scoring functions can be categorized into three types:

- Forcefield-based scoring functions
- Empirical scoring functions
- Knowledge-based scoring functions

Forcefield-based scoring functions

Most scoring functions are physics-based molecular mechanics force fields that estimate the energy of the pose within the binding site. The various contributions to binding can be written as an additive equation:

$$E = W_{VDW}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}+\frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}p(\theta)\left(\frac{C_{ij}}{r_{ij}^{12}}+\frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec}\sum_{i.j}\frac{q_i q_j}{r_{ij}} + W_{sol}\sum_{i.j}(S_i V_j + S_j V_i)e^{\left(-r_{ij}^2/2\sigma^2\right)}$$

The components in this example consist of van der Waals (W_{VDW}) and electrostatic (W_{elec}) interactions, solvent effects (W_{sol}) and hydrogen bonding interactions (W_{hbond}).

The **van der Waals interaction term** describes the non-bonded interactions between non-polar atoms such as hydrophobic interactions. It is a standard Lennard-Jones potential function and the W_{VDW} is a scaling factor to bring the van der Waals contributions on the same scale as the other parameters.

Hydrogen bond interactions are modelled in a similar way as the van der Waals interactions, with the exception that a 12-10 Lennard-Jones function is used with different parameters, and that a penalty function $p(\theta)$ has been added to account for deviations of the hydrogen bond from ideal geometry (Figure 83).

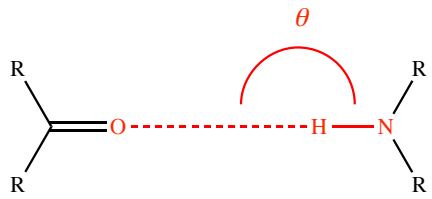

Figure 83. Illustration of the θ angle to calculate hydrogen bond geometry. Shown is a hydrogen bond between the O and H-N. In a hydrogen bond of optimal strength, this O-H-N angle should be close to 180°. Deviations from this optimal value results in weaker hydrogen bonding interactions.

Electrostatic interactions are normally represented as a constant W_{elec} term multiplied by the product of the partial charges and divided by the atomic distances.

Finally, **solvent effects** can be modelled in a variety of ways; the method which is shown in the equation above is based on the surfaces of ligand and receptor atoms (S_i and S_j), scaled by a solvent parameter (V_i and V_j) and the distance between the atoms (r_{ij}).

Empirical scoring functions

Empirical scoring functions are based on counting the number of various types of interactions between the two binding partners. Counting may be based on the number of ligand and receptor atoms in contact with each other or by calculating the change in solvent accessible surface area (ΔSASA) in the complex compared to the uncomplexed ligand and protein. The coefficients of the scoring function are usually derived by fitting using multiple linear regression methods, in which experimental structures and corresponding binding affinities are used as model system. These interactions terms of the function may include for example:

$$\Delta G = f_{hbonds}\Delta G_{hbonds} + f_{polar-apolar}\Delta G_{lpolar-apolar} + f_{nrot}\Delta G_{nrot} + f_{apolar-apolar}\Delta G_{apolar-apolar}$$

with f_{hbonds} a function representing the number of hydrogen bonds and ΔG_{hbonds} the corresponding parameter resulting from the least-squares fit, $f_{polar\text{-}apolar}$ and $\Delta G_{polar\text{-}apolar}$ the function and parameter accounting for the unfavourable hydrophobic-hydrophilic contacts between receptor and ligand, $f_{apolar\text{-}apolar}$ and $\Delta G_{apolar\text{-}apolar}$ the function and parameter accounting for the favourable hydrophobic-hydrophobic contacts, and f_{nrot} and ΔG_{nrot} the function and parameter representing the number of rotational bonds on the ligand (accounting for the loss of entropy upon ligand binding).

Knowledge-based scoring functions

An third scoring function approach is to derive a knowledge-based statistical potential for interactions from a large database of protein-ligand complexes, such as the Protein Data Bank (https://www.rcsb.org) or the Cambridge Structural Database (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/), and evaluate the fit of the pose according to this inferred potential. Knowledge-based scoring functions are based on statistical observations of intermolecular close contacts in large 3D databases which are used to derive 'potentials of mean force'. This method is founded on the assumption that close intermolecular interactions between certain types of atoms or functional groups that occur more frequently than one would expect by a random distribution are likely to be energetically favourable and therefore contribute favourably to binding affinity.

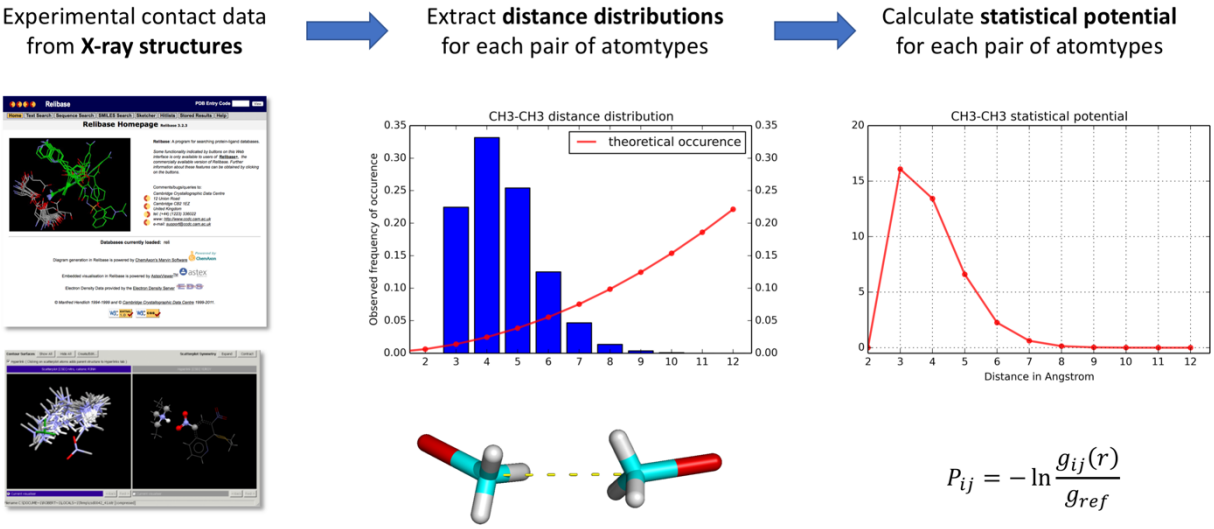

Figure 84. Development of a knowledge-based scoring function explained for the methyl-methyl interaction. Knowledge about the particular interaction type is initially extracted from structural databases from the number of methyl-methyl interactions is counted and also the corresponding geometrical features, such as the carbon-carbon distance. This experimental methyl-methyl distance distribution is subsequently compared to a theoretical distribution, i.e. 'what should be the distribution from a pure theoretical point of view, without the presence of favorable or unfavorable interaction terms'. Dividing the observed distribution g_{ij} by the reference distribution g_{ref} yields the potential of mean force that describes the potential between two methyl groups as a function of distance.

Scope of the scoring functions

Docking experiments are used for pose prediction or for compound selection. In the case of **pose prediction**, the question that needs to be answered is 'given a ligand and a protein, how does the ligand bind to the protein?'. Put in other words, the docking program needs to predict a valid binding conformation but the corresponding predicted binding affinity should not be accurate. Emphasis in this case is on predicting the correct binding pose rather than on predicting binding affinity values. However, in the case of **compound selection**, the scope is different. In this situation, one often wants to select from a large compound database those compounds that will bind to the protein under study, and therefore emphasis is on predicting correct binding affinities rather than correct binding poses.

Forcefield-based scoring functions are most suitable for correct pose prediction but less suited for compound selection. The reason for this is that the value of the scores generated by a forcefield-based function are dependent on the ligand type and its size, and therefore more difficult to compare between different ligands. However, the calculated scores are accurate enough to compare different docking poses of the same ligand, and for that reason forcefield-based scoring functions are suitable for pose prediction questions.

Empirical scoring functions are better suited for selection of compounds from a database screen. First of all, because of their simplicity, these scoring functions are fast to calculate and therefore suited to screen millions of compounds, and secondly, because these functions were derived by least-square fitting against experimental binding affinities, the resulting scores are also less depending on the ligand type and size and therefore

comparable between different ligands. The simplicity of the function has a drawback in the sense that the results are less sensitive to small variation in the docking poses, hence less suitable for pose prediction questions.

Knowledge-based scoring functions have been shown to be useful in both pose prediction as well as database screening cases. These functions are accurate enough to pick up small variations in the actual binding poses, and the score values are rather independent on the ligand type and size, hence also applicable to database screening. The main issue with these kind of scoring functions is the fact that there are a large number of structures from X-ray crystallography for complexes between proteins and high affinity ligands, but comparatively fewer for low affinity ligands as the later complexes tend to be less stable and therefore more difficult to crystallize. Scoring functions trained with this high affinity data can therefore dock high affinity ligands correctly, but as a consequence they will also give plausible docked conformations for ligands that do not bind. This gives a large number of false positive hits, i.e., ligands predicted to bind to the protein that actually don't when tested experimentally.