1

Mixing Pharmacophore Modeling and Classical QSAR Analysis as Powerful Tool for Lead Discovery

Mutasem O. Taha
Dept. of Pharmaceutical Sciences, Faculty of Pharmacy, University of Jordan
Jordan

1. Introduction

Discovery of new bioactive leads for subsequent optimization into drugs is both time consuming and expensive process. Two main approaches are currently available for lead discovery, namely, high throughput (*in vitro*) screening and computer-aided virtual (*in silico*) screening. Normally, *in silico* techniques are implemented as pre-filters to enrich the success rates of high throughput screening campaigns.

Computer-aided lead discovery techniques can be divided into two main methodologies: (1) Structure-based and (2) ligand-based methods. Structure-based methods depend on the availability of three-dimensional (3D) structure for the targeted biomolecule (enzyme of receptor). The target structure is normally employed as template to design sterically and electronically complementary ligands. On the other hand, ligand-based methods rely on assessing physicochemical and structural similarities among potent ligands and try to discern ligands' structural features responsible for high affinities from those responsible for poor affinities. In other words, ligand-based methods rely completely on ligand factors to assess ligand-receptor affinities.

Structure-based drug design can be divided into two major methodologies: de novo and docking-based design. De novo design involves the use of algorithms that construct virtual ligands inside binding pockets.[1-3] On the other hand, docking involves fitting virtual ligands, usually from large virtual libraries, into targeted binding sites employing computer algorithms that rely on force fields to calculate attractive and repulsive interactions within virtual ligand-protein complexes.[1-4]

The 3D structures of targeted receptors or enzymes are generally obtained via X-ray crystallographic scattering, nuclear magnetic resonance (NMR) or homology modeling.[1-4] However, reliance on crystallographic structures represents a major problem for structure-based design. Crystallographic structures are limited by inadequate resolution[5] and crystallization-related artifacts of the ligand-protein complex.[6-8] Moreover, crystallographic structures generally ignore structural heterogeneity related to protein anisotropic motion and discrete conformational substrates.[9]

Moreover, molecular docking, which is basically a conformational sampling procedure in which various docked conformations are explored to identify the right one, can be very

challenging problem given the degree of conformational flexibility at the ligand-macromolecular level.[10-12] Although docking programs employ diverse methodologies to evaluate different ligand conformations within binding pockets,[13-23] conformational sampling must be guided by scoring function(s) to evaluate the fitness between the protein and the ligand.[4,24-29] The final docked conformations are selected according to their scores. Unfortunately, the sheer complexity of the underlying ligand-receptor molecular interactions extremely complicate free energy calculations and undermine the ability of scoring functions to evaluate binding free energies correctly in order to rank different potential ligand–receptor complexes.[1,3,5-8,10,30,31,39]

In addition to deciding the optimal docking/scoring combination for a particular docking problem, the molecular modeler must decide whether to leave crystallographically explicit water molecules in the binding site or not prior to ligand docking.[32-37] Furthermore, the fact that crystallographic structures lack information on hydrogen atoms means that it should be appropriately assumed whether ionizable moieties embedded within the binding site exist in their ionized form or not.[36,38]

Additional to the previous problems, use of single protein conformation for designing new ligands ignores important dynamic aspects of protein-ligand binding. In particular, the "induced fit" effects are ignored.[40,41] Unfortunately, all current computational models directed towards assessing the flexibility of macromolecular binding sites (e.g., soft receptors[42,43]; few critical rotable degrees of freedom in the receptor binding site[43-48]; systematic conformer searches of amino acids' side chains at the binding site[49]; molecular dynamics and free energy calculations conducted on flexible enzyme[50,51]; use of multiple crystallographic receptor structures[43,52]) suffer from two major drawbacks. Firstly, their computational cost, which reduces their effectiveness in virtual screening and fast docking, and secondly, their complete reliance on crystallographic structures.

The drawbacks of structure-based methods prompted us to introduce an interesting and novel ligand-based approach as a tool for characterizing binding sites' flexibilities. This approach ignores the protein template and focuses completely on the ligand side. It is carried out over two subsequent stages. Firstly, the pharmacophoric space of the targeted enzyme is extensively explored utilizing the three-dimensional Quantitative Structure Activity Relationship (3D-QSAR) software program CATALYST. The resulting binding models (hundreds) are then in allowed to compete within the context of classical quantitative structure-activity relationship analyses (QSAR) employing genetic algorithm (GA) and multiple linear regression (MLR) analyses. This process selects optimal combination of orthogonal pharmacophores that best explain the observed bioactivities, i.e., best possible QSAR model. Such combination of binding pharmacophores should correspond to accessible binding modes available for ligands within a particular binding pocket.

We previously reported the successful use of this combination to probe the induced fit flexibilities of activated factor X[53] and towards the discovery of new inhibitory leads against glycogen synthase kinase-3β,[54] bacterial MurF,[55] protein tyrosine phosphatase,[56] DPP IV,[57] hormone sensitive lipase,[58] β-secretase,[59] influenza neuraminidase,[60] cholesteryl ester transfer protein,[61] cycline dependent kinase,[62] Heat shock protein,[63] estrogen receptor β,[64] β-D-Glucosidase,[65] and β-D-Galactosidase.[66]

The author intends in this chapter to discuss the basic theoretical principles of this successful ligand-based approach and to provide interested audiences with experimental details related to this approach.

The modeling process of this approach can be divided in the following steps:

2. Data mining and conformational coverage

Firstly, the literature is extensively surveyed to identify as many reported structurally diverse ligands against the selected target as possible. The collected compounds must satisfy two important prerequisites: (i) they should be all bioassayed by a single procedure. Consistency in bioassay is a major requirement for QSAR modeling as it is not possible to model bioactivity data generated via more than one bioassay procedure. (ii) They must exhibit wide bioactivity range, i.e., over > 4 logarthmic cycles.

Initially, the 2D structures of the inhibitors are imported into the modeling package (CATALYST) and converted automatically into plausible 3D single conformer representations and energy minimized to the closest local minimum. The resulting single conformer 3D structures are normally used as starting points for conformational analysis for pharmacophore modeling and in the determination of various molecular descriptors for QSAR modeling.

The conformational space of each ligand is extensively sampled usually utilizing the poling algorithm employed within the CONFIRM module of CATALYST.[67] Efficient conformational coverage guarantees minimum conformation-related noise during pharmacophore generation and validation stages because pharmacophore generation and pharmacophore-based search procedures are known for their sensitivity to inadequate conformational sampling within the training compounds.[60]

The logarithm of measured IC_{50}, EC_{50} or Ki values are used in pharmacophore modeling and QSAR analysis, thus correlating the data linear to the free energy change.

2.1 Exploration of pharmacophoric space

2.1.1 The algorithm

Normally we implement the HYPOGEN module within CATALYST package to explore the pharmacophoric space of different ligands. CATALYST-HYPOGEN enables automatic pharmacophore construction by using a collection of at least 16 molecules with bioactivities spanning over 3.5 orders of magnitude.[67]

CATALYST-HYPOGEN models drug-receptor interaction using information derived from the ligand structure. It identifies a 3D array of a maximum of five chemical features common to active training molecules, which provides a relative alignment for each input molecule consistent with their binding to a proposed common receptor site. The chemical features considered can be hydrogen bond donors and acceptors (HBDs and HBAs), aliphatic and aromatic hydrophobes (Hbic), positive and negative ionizable (PosIon and NegIon) groups and aromatic planes (RingArom). CATALYST pharmacophores have been used as 3D queries for database searching and in 3D-QSAR studies.[54-66]

Although pharmacophore modeling employing HYPOGEN has been heavily reviewed in the literature,[68-76] a brief discussion of this algorithm is provided herein to allow better readability of the chapter.

HYPOGEN pharmacophore exploration proceeds through three successive phases: the constructive phase, subtractive phase and optimization phase.[67-76]

During the constructive phase, CATALYST generates common conformational alignments among the most-active training compounds. Only molecular alignments based on a maximum of five chemical features are considered. The program identifies a particular compound as being within the most active category if it satisfies equation (1).[73]

$$(\text{MAct} \times \text{UncMAct}) - (\text{Act} / \text{UncAct}) > 0.0 \quad (1)$$

Where "MAct" is the activity of the most active compound in the training set, "Unc" is the uncertainty of the compounds and "Act" is the activity of the training compounds under question. However, if there are more than eight most-active inhibitors, only the top eight are used.

In the subsequent subtractive phase, CATALYST eliminates some hypotheses that fit inactive training compounds. A particular training compound is defined as being inactive if it satisfies equation (2): [67-76]

$$\text{Log (Act)} - \text{log (MAct)} > 3.5 \quad (2)$$

However, in the optimization phase, CATALYST applies fine perturbations in the form of vectored feature rotation, adding new feature and/or removing a feature, to selected hypotheses that survived the subtractive phase, in an attempt to find new models of enhanced bioactivity/mapping correlation, i.e., improved 3D-QSAR properties.[67-76]

Eventually, CATALYST selects the highest-ranking models (10 by default) and presents them as the optimal pharmacophore hypotheses resulting from the particular automatic modeling run.

2.1.2 Selection of training subsets

The fact that pharmacophore modeling requires limited number of carefully selected training compounds (from 16-45 compounds only)[67-76] that exhibit bioactivity variations attributable solely to the presence or absence of pharmacophoric features, i.e., not due to steric or electronic factors, makes it impossible to explore the pharmacophore space of large training sets in one shot (e.g., we normally collect more that 100 compounds), partly because CATALYST-HYPOGEN is not suited to handle large number of compounds and partly because pharmacophore modeling is generally confused by electronic and steric bioactivity modifying factors commonly encountered in SAR data. This dilemma prompted us to break compound lists into smaller training subsets compatible with pharmacophore modeling, i.e., of bioactivity variations attributable solely to the presence or absence of pharmacophoric features. Nevertheless, the basic problem in this approach is to identify a particular training set capable of representing the whole list of collected compounds. This problem can be very significant in cases of large SAR lists. We found that the best way to solve this problem is by exploring the pharmacophoric space of several carefully selected training

subsets, i.e., from the whole list of collected compounds, followed by allowing the resulting pharmacophores to compete within the context of genetic function approximation-based quantitative structure-activity relationship (GFA-QSAR) analysis such that the best pharmacophore(s) capable of explaining bioactivity variations across the whole list of collected compounds is(are) selected. However, since pharmacophore models fail in explaining electronic and steric bioactivity-modulating effects, the GFA-QSAR process should be allowed to select other 2D physicochemical descriptors to complement the selected pharmacophore(s) (see below).

The training compounds in these subsets are selected in such away to guarantee maximal 3D diversity and continuous bioactivity spread over more than 3.5 logarithmic cycles. Moreover, training subsets are selected in such a way that their member compounds share certain apparent 3D SAR rules (by visual evaluation).

We usually give special emphasis to the 3D diversity of the most active compounds in each training subset because of their significant influence on the extent of the evaluated pharmacophoric space during the constructive phase of HYPOGEN algorithm. However, it must be mentioned that not all collected compounds are incorporated in the pharmacophore training subsets, in fact, compounds that exhibit limited diversity or significant bioactivity-modifying steric or electronic influences are excluded from the training subsets.

2.1.3 Modeling boundaries

HYPOGEN implements an optimization algorithm that evaluates large number of potential binding models for a particular target through fine perturbations to hypotheses that survived the constructive and subtractive phases of the modeling algorithm.[67-76] The extent of the evaluated pharmacophoric space is reflected by the configuration (Config.) cost calculated for each modeling run. It is generally recommended that the Config. cost of any HYPOGEN run not to exceed 17 (corresponding to 2^{17} hypotheses to be assessed by HYPOGEN) to guarantee thorough analysis of all models.[71-73] The size of the investigated pharmacophoric space is a function of training compounds, selected input chemical features and other CATALYST control parameters.[67-76]

We envisaged that restricting the size of explored pharmacophoric space should improve the efficiency of optimization via allowing efficient assessment of limited number of pharmacophoric models. On the other hand, extreme restrictions imposed on the evaluated pharmacophoric space might reduce the possibility of discovering optimal binding hypotheses, as they might occur outside the "boundaries" of the evaluated space.

Therefore, we normally explore the pharmacophoric space of targeted ligands under reasonably imposed "boundaries" through numerous HYPOGEN runs and employing several carefully selected training subsets.

Guided by our rationally restricted pharmacophoric exploration concept, we usually restrict HYPOGEN to explore pharmacophoric models incorporating limited number of features, e.g., from zero to one negative NegIon, or PosIon features or from zero to three HBA, Hbic, and RingArom features instead of the default range of zero to five. Furthermore, we normally instructed HYPOGEN to explore only 4- and 5-featured pharmacophores, i.e., ignore models of lesser number of features in order to further narrow the investigated

pharmacophoric space and to better represent the diverse interactions between known ligands and binding pockets. In fact, three- and two-featured pharmacophores are rather promiscuous as 3D search queries and not adequate descriptions of ligand-receptor binding.

2.1.4 Assessment of generated pharmacophore models

When generating hypotheses, CATALYST attempts to minimize a cost function consisting of three terms: Weight cost, Error cost and Configuration cost.[67-76] Weight cost is a value that increases as the feature weight in a model deviates from an ideal value of 2. The deviation between the estimated activities of the training set and their experimentally determined values adds to the error cost. The activity of any compound can be estimated from a particular hypothesis through equation (3).[73]

$$\text{Log (Estimated Activity)} = I + \text{Fit} \tag{3}$$

Where, I = the intercept of the regression line obtained by plotting the log of the biological activity of the training set compounds against the Fit values of the training compounds. The Fit value for any compound is obtained automatically employing equation (4).[73]

$$\text{Fit} = \Sigma \text{ mapped hypothesis features} \times W\,[1 - \Sigma\,(\text{disp}/\text{tol})^2] \tag{4}$$

Where, Σ mapped hypothesis features represents the number of pharmacophore features that successfully superimpose (i.e., map or overlap with) corresponding chemical moieties within the fitted compound, W is the weight of the corresponding hypothesis feature spheres. This value is fixed to 1.0 in CATALYST-generated models. disp is the distance between the center of a particular pharmacophoric sphere (feature centroid) and the center of the corresponding superimposed chemical moiety of the fitted compound; tol is the radius of the pharmacophoric feature sphere (known as Tolerance, equals to 1.6 Å by default). $\Sigma\,(\text{disp}/\text{tol})^2$ is the summation of $(\text{disp}/\text{tol})^2$ values for all pharmacophoric features that successfully superimpose corresponding chemical functionalities in the fitted compound.[67-76]

The third cost term, i.e., the configuration cost, penalizes the complexity of the hypothesis. This is a fixed cost, which is equal to the entropy of the hypothesis space. The more the numbers of features (a maximum of five) in a generated hypothesis, the higher is the entropy with subsequent increase in this cost. The overall cost (total cost) of a hypothesis is calculated by summing over the three cost factors. However, error cost is the main contributor to the total cost.

CATALYST also calculates the cost of the null hypothesis, which presumes that there is no relationship in the data and that experimental activities are normally distributed about their mean. Accordingly, the greater the difference from the null hypothesis cost, the more likely that the hypothesis does not reflect a chance correlation. In a successful automatic modeling run, CATALYST ranks the generated models according to their total costs.[67-76]

An additional approach to assess the quality of CATALYST-HYPOGEN pharmacophores is to cross-validate them using the Cat-Scramble module implemented in CATALYST. This validation procedure is based on Fisher's randomization test.[43] In this validation test, a 95% confidence level was selected, which instruct CATALYST to generate 19 random

spreadsheets by the Cat-Scramble command. Subsequently, CATALYST-HYPOGEN is challenged to use these random spreadsheets to generate hypotheses using exactly the same features and parameters used in generating the initial unscrambled hypotheses.[67] Success in generating pharmacophores of comparable cost criteria to those produced by the original unscrambled data reduces the confidence in the training compounds and the unscrambled original pharmacophore models.

Eventually, the top 10 binding hypotheses (i.e., pharmacophores) from each automatic HYPOGEN run are automatically ranked according to their corresponding "total cost" values and presented as output of the HYPOGEN run.

2.2 Clustering of successful pharmacophore hypotheses

Because the number of generated pharmacophores during our pharmacophore exploration step is usually large (> 60 model) and they usually share several 3D features and properties (cost criteria, Cat.scramble confidence, etc …), we normally cluster the resulting models into limited number of groups (10-30) utilizing the hierarchical average linkage method available in CATALYST. The highest-ranking hypothesis within each cluster (i.e., of lowest cost or highest correlation with bioactivity of the whole collected list) is selected to represent the corresponding cluster in subsequent QSAR modeling.

Clustering aims at avoiding overloading genetic function approximation-multiple linear regression (GFA-MLR), implemented during QSAR modeling, with numerous independent variables, which may allow the emergence of less-than-optimal regression models.

2.3 QSAR modeling

Pharmacophoric hypotheses are important tools in drug design and discovery as they provide excellent insights into ligand-macromolecule recognition and they can be used to mine for new biologically interesting scaffolds. However, their predictive value as 3D-QSAR models is usually limited by steric shielding and bioactivity-enhancing or -reducing auxiliary groups.[76] This point combined with the fact that pharmacophore exploration usually furnish several binding hypotheses of comparable success criteria and 3D features prompt us to use classical QSAR analysis to search for optimal combination of pharmacophore(s) and other 2D descriptors capable of explaining bioactivity variation across the whole list of collected inhibitors. We normally employ genetic function approximation and multiple linear regression QSAR (GFA-MLR-QSAR) analysis to search for an optimal QSAR equation(s) using the logarithm of measured $1/IC_{50}$ or 1/Ki values are as dependent variables (thus correlating the data linear to the free energy change).

GFA-MLR-QSAR selects optimal descriptor combinations based on the Darwinian concept of genetic evolution whereby the statistical criteria of regression models from different descriptor combinations (chromosomes) are employed as fitness criteria.[77] GFA-MLR-QSAR analysis is employed to explore various combinations of pharmacophores and other structural descriptors and to evaluate their statistical properties as predictive QSAR models.

Representative pharmacophore hypotheses (selected during the clustering stage) are fitted against all collected ligands and their fit values (determined by equation 4) are enrolled, together with other 2D and 1D structural descriptors, as independent variables (genes) in a cycle of GFA-MLR-QSAR analysis over thousands of iterations.[77]

Other structural descriptors include various simple and valence connectivity indices, electro-topological state indices and other molecular descriptors (e.g., logarithm of partition coefficient, polarizability, dipole moment, molecular volume, molecular weight, molecular surface area, etc.).[77]

However, to assess the predictive power of the optimal QSAR models on external set of inhibitors, we usually randomly select around 20% of the collected ligands and employ them as external testing molecules for validating optimal QSAR model(s) (r^2_{PRESS}). Moreover, all QSAR models are cross-validated automatically using the leave-one-out cross-validation.[77]

Emergence of two or more orthogonal pharmacophoric models in the optimal QSAR model suggests the existence of complementary two or more corresponding binding modes accessible to ligands within the binding pocket of target protein, i.e., one of the pharmacophores can optimally explain the bioactivities of some training inhibitors, while the others explain the remaining inhibitors. Such conclusions were reached about the binding pockets of several targets, e.g., factor Xa, GSK-3β, and Mur F.[57-63]

2.4 Final validation of optimal QSAR model and associated pharmacophores

To establish the validity of optimal GFA-selected QSAR model and associated pharmacophore(s), we normally implement two validation methods: (1) Receiver-Operating Characteristic (ROC) curve analysis, and (2) Comparing QSAR-selected pharmacophore(s) with the corresponding binding site, however, this is only done upon having available crystallographic structure of the targeted receptor.

2.4.1 Receiver Operating Characteristic (ROC) curve analysis

In ROC analysis, the ability of a particular pharmacophore model to correctly classify a list of compounds as actives or inactives is indicated by the area under the curve (AUC) of the corresponding ROC as well as other parameters, namely, overall accuracy, overall specificity, overall true positive rate and overall false negative rate.[78-79]

The testing list for ROC analyses are usually prepared as described by Verdonk and co-workers.[78] Briefly, decoy compounds are selected based on three basic one-dimensional (1D) properties that allow the assessment of distance (D) between two molecules (e.g., i and j): (1) the number of hydrogen-bond donors (NumHBD); (2) number of hydrogen-bond acceptors (NumHBA) and (3) count of nonpolar atoms (NP, defined as the summation of Cl, F, Br, I, S and C atoms in a particular molecule). For each active compound in the test set, the distance to the nearest other active compound is assessed by their Euclidean Distance (Equation 5):

$$D(i,j)=\sqrt{(NumHBD_i - NumHBD_j)^2 + (NumHBA_i - NumHBA_j)^2 + (NP_i - NP_j)^2} \quad (5)$$

The minimum distances are then averaged over all active compounds (Dmin). Subsequently, for each active compound in the test set, around 40 decoys are randomly chosen from the ZINC database.[80] The decoys are selected in such a way that they did not exceed Dmin distance from their corresponding active compound.

To diversify active members in the list, we exclude any active compound having zero distance ($D(i,j)$) from other active compound(s) in the test set.

The test set is then screened by each particular pharmacophore employing the "Best flexible search" option implemented in CATALYST, while the conformational spaces of the compounds are usually generated employing the "Fast conformation generation option" implemented in CATALYST. Compounds missing one or more features were discarded from the hit list. *In-silico* hits were scored employing their fit values as calculated by equation (4).

The ROC curve analysis describes the sensitivity (Se or true positive rate, equation 6) for any possible change in the number of selected compounds (n) as a function of (1-Sp). Sp is defined as specificity or true negative rate (equation 7).[79,81]

$$Se = \frac{Number\ of\ Selected\ Actives}{Total\ Number\ of\ Actives} = \frac{TP}{TP + FN} \tag{6}$$

$$Sp = \frac{Number\ of\ Discarded\ Inactives}{Total\ Number\ of\ Inactives} = \frac{TN}{TN + FP} \tag{7}$$

where, TP is the number of active compounds captured by the virtual screening method (true positives), FN is the number of active compounds discarded by the virtual screening method, TN is the number of discarded decoys (presumably inactives), while FP is the number of captured decoys (presumably inactive).[79,81]

If all molecules scored by a virtual screening (VS) protocol with sufficient discriminatory power are ranked according to their score (i.e., fit values), starting with the best-scored molecule and ending with the molecule that got the lowest score, most of the actives will have a higher score than the decoys. Since some of the actives will be scored lower than decoys, an overlap between the distribution of active molecules and decoys will occur, which will lead to the prediction of false positives and false negatives.[79,81] The selection of one score value as a threshold strongly influences the ratio of actives to decoys and therefore the validation of a VS method. The ROC curve method avoids the selection of a threshold by considering all Se and Sp pairs for each score threshold.[79,81] A ROC curve is plotted by setting the score of the active molecule as the first threshold. Afterwards, the number of decoys within this cutoff is counted and the corresponding Se and Sp pair is calculated. This calculation is repeated for the active molecule with the second highest score and so forth, until the scores of all actives are considered as selection thresholds.

The ROC curve representing ideal distributions, where no overlap between the scores of active molecules and decoys exists, proceeds from the origin to the upper-left corner until all the actives are retrieved and Se reaches the value of 1. In contrast to that, the ROC curve for a set of actives and decoys with randomly distributed scores tends towards the Se = 1-Sp line asymptotically with increasing number of actives and decoys.[79,81] The success of a particular virtual screening workflow can be judged from the following criteria:

1. Area under the ROC curve (AUC).[79,81] In an optimal ROC curve an AUC value of 1 is obtained; however, random distributions cause an AUC value of 0.5. Virtual screening that performs better than a random discrimination of actives and decoys retrieve an

AUC value between 0.5 and 1, whereas an AUC value lower than 0.5 represents the unfavorable case of a virtual screening method that has a higher probability to assign the best scores to decoys than to actives.[79,81]

2. Overall Accuracy (ACC): describes the percentage of correctly classified molecules by the screening protocol (equation 8). Testing compounds are assigned a binary score value of zero (compound not captured) or one (compound captured).[79,81]

$$ACC = \frac{TP + TN}{N} = \frac{A}{N}.Se + \left(1 - \frac{A}{N}\right).Sp \quad (8)$$

where, N is the total number of compounds in the testing database, A is the number of true actives in the testing database.

3. Overall specificity (SPC): describes the percentage of discarded inactives by the particular virtual screening workflow. Inactive test compounds are assigned a binary score value of zero (compound not captured) or one (compound captured) regardless to their individual fit values.[79,81]
4. Overall True Positive Rate (TPR or overall sensitivity): describes the fraction percentage of captured actives from the total number of actives. Active test compounds are assigned a binary score value of zero (compound not captured) or one (compound captured) regardless to their individual fit values.[79,81]
5. Overall False Negative Rate (FNR or overall percentage of discarded actives): describes the fraction percentage of active compounds discarded by the virtual screening method. Discarded active test compounds are assigned a binary score value of zero (compound not captured) or one (compound captured) regardless to their individual fit values.[79,81]

2.5 *In Silico* screening

Eventually, optimal QSAR-selected pharmacophores are employed as 3D search queries against several electronic multiconformer structural databases (e.g. NCI 238,819 structures) using the "Best Flexible Database Search" option implemented within CATALYST. Compounds that have their chemical groups spatially overlap (map) with corresponding features of the particular pharmacophoric model are captured as hits. Hits are normally filtered based on Lipinski's and Veber's rules.[82,83] Surviving hits are then fitted against QSAR-selected pharmacophores and their fit values, together with other relevant molecular descriptors, are substituted in optimal QSAR equation to predict their bioactivities. The highest-ranking available hits are evaluated *in vitro*.

Usually, the acquired hits are screened at 10 μM concentrations, subsequently; compounds of significant bioactivities at 10 μM are further assessed to determine their IC_{50} values.

It remains to be mentioned that although QSAR predictions are rather accurate with some hit compounds, experimental IC_{50} values of other hits differ significantly from QSAR predictions. These errors appear are usually related to structural differences between training compounds used in QSAR and pharmacophore modeling compared to hit molecules. This discrepancy seems to limit the extrapolatory potential of the QSAR equation.

3. Conclusions

This chapter summarizes an interesting novel approach for the discovery of new bioactive leads by implementing a sequential process of pharmacophore modeling and QSAR analysis. This approach has been used for the discovery of potent inhibitors against at least a dozen enzymes and receptors.

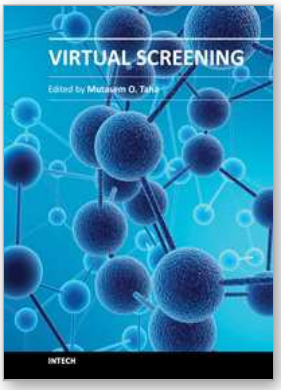
4. References

[1] Song, C. M.; Lim, S. J.; Tong, J. C. R. *Brief Bioinform.* 2009, 10, 579–591.
[2] Menikarachchi, L. C.; Gascon, J. A. *Curr. Top. Med. Chem.* 2010, 10, 46–54.
[3] Jorgensen, W. L. *Acc. Chem. Res.* 2009, 42, 724–733.
[4] Hecht, D.; Fogel, G. B. *Curr. Comput.-Aided Drug Des.* 2009, 5, 56–68.
[5] Beeley, N. R. A.; Sage, C. *Targets* 2003, 2, 19–25.
[6] Klebe, G. *Drug Discovery Today* 2006, 11, 580–594.
[7] Steuber, H.; Zentgraf, M.; Gerlach, C.; Sotriffer, C. A.; Heine, A; Klebe, G. *J. Mol. Biol.* 2006, 363, 174–187.
[8] Stubbs, M. T.; Reyda, S.; Dullweber, F.; Moller, M.; Klebe, G.; Dorsch, D.; Mederski, W.; Wurziger, H. *ChemBioChem* 2002, 3, 246–249.
[9] DePristo, M. A.; de Bakker, P. I. W.; Blundell, T. L. *Structure* 2004, 12, 831–838.
[10] Morris, G. M.; Olson, A. J.; Goodsell, D. S. *Princ. Med. Chem*. 2000, 8, 31–48.
[11] Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. *J. Med. Chem.* 2004, 47, 558–565.
[12] Beier, C.; Zacharias, M. *Expert Opin. Drug Dis.* 2010, 5, 347–359.
[13] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol*. 1996, 261, 470–489.
[14] Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *J. Comput. Aid Mol. Des.* 2001, 15, 411–428.
[15] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* 1997, 267, 727–748.
[16] Vaque, M.; Ardrevol, A.; Blade, C.; Salvado, M. J; Blay, M.; Fernandez-Larrea, J.; Arola, L.; Pujadas, G. *Curr. Pharm. Anal.* 2008, 4, 1–19.
[17] Cosconati, S.; Forli, S.; Perryman, A. L.; Harris, R.; Goodsell, D. S.; Olson, A. J. *Expert Opin. Drug Dis.* 2010, 5, 597–607.
[18] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J. Comput. Chem.* 1998, 19, 1639–1662.
[19] Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. *J. Med. Chem.* 2004, 47, 1750–1759.
[20] CERIUS2 LigandFit, version 4.10; Accelrys, Inc.: San Diego, 2000.
[21] FRED, version 2.1; OpenEye Scientific Software: Santa Fe, NM, 2006.
[22] Diller, D. J.; Merz, K. M. *Proteins* 2001, 43, 113–124.
[23] Rao, S. N.; Head, M. S.; Kulkarni, A.; LaLonde, J. M. *J. Chem. Inf. Model.* 2007, 47, 2159–2171.
[24] Bissantz, C.; Folkers, G.; Rognan, D. *J. Med. Chem.* 2000, 43, 4759–4767.
[25] Gao, W. R.; Lai, Y. L. *J. Mol. Model.* 1998, 4, 379–394.
[26] Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. *J. Mol. Graphics Modell*. 2005, 23, 395–407.
[27] Velec, H. F. G.; Gohlke, H.; Klebe, G. *J. Med. Chem.* 2005, 48, 6296–6303.

[28] Jain, A. N. *Curr. Protein Pept. Sci.* 2006, 7, 407–20.
[29] Rajamani, R.; Good, A. C. *Curr. Opin. Drug Discovery Dev.* 2007, 10, 308–15.
[30] Tame, J. R. H. *J. Comput.-Aided Mol. Des.* 1999, 13, 99–108.
[31] Kollman, P. *Chem. Rev.* 1993, 93, 2395–2417.
[32] Homans, S. W. *Drug Discovery Today* 2007, 12, 534–539.
[33] Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* 1995, 9, 500–512.
[34] Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* 1995, 9, 513–520.
[35] Poornima, C. S.; Dean, P. M. *J. Comput.-Aided Mol. Des.* 1995, 9, 521–531.
[36] Koehler, K. F.; Rao, S. N.; Snyder, J. P. Modeling drug-receptor interactions. In Guidebook on Molecular Modeling in Drug Design; Cohen, N. C., Ed.; Academic Press; San Diego, 1996; pp 235-336.
[37] Pastor, M.; Cruciani, G.; Watson, K. *J. Med. Chem.* 1997, 40, 4089–4102.
[38] Silverman, R. A. The Organic Chemistry of Drug Design and Drug Action; Academic Press: San Diego, 1991, pp 62-65.
[39] Krissinel, E. *J. Comput. Chem.* 2009, 31, 133–143.
[40] D. E. Koshland, *Proc. Natl Acad. Sci. USA* 1958, 44, 98-104.
[41] W.L. Jorgensen, Science 254 (1991) 951-955.
[42] Cohen N. C., Guidebook on Molecular Modeling in Drug Design, Academic Press, UK, 1996.
[43] M. L. Teodoro, L. E. Kavraki, *Curr. Pharm. Design* 2003, 9, 1635-1648.
[44] R. L. Dunbrack, M. Karplus, *J. Mol. Biol.* 1993, 230, 543-574.
[45] G. Vriend, C. Sander, P. F. W. Stouten, *Prot. Eng*. 1994, 7 23-29.
[46] H. Shrauber, F. Eisenhaber, P. Argos, *J. Mol. Biol.* 1993, 230 592-612.
[47] A. R. Leach, I. D. Kuntz, *J. Comput. Chem.* 1992, 13, 730-748.
[48] R. Leach, *J. Mol. Biol.* 1994, 235, 345-356.
[49] F. Eisenmenger, P. Argos, R. Abagyan, *J. Mol. Biol.* 1993, 231 849-860.
[50] P. Kollman, Curr. Opin. Struct. Biol. 4 (1994) 240-245.
[51] McCammon, J. A., Harvey, S. C., Dynamics of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1987.
[52] R. M. A. Knegtel, I. D. Kuntz, C. M. Oshiro, *J. Mol. Biol.* 1997, 266 , 424-440.
[53] Taha, Mutasem O.; Qandil, Amjad M.; Zaki, Dhia D.; Murad A. AlDamen. *Eur. J. Med. Chem.* 2005, 40, 701-727.
[54] Taha, M.O.; Bustanji, Y.; Al-Ghussein, M.A.S.; Mohammad, M.; Zalloum, H.; Al-Masri, I.M.; Atallah, N. *J. Med. Chem.* 2008, 51, 2062-2077.
[55] Taha, M.O.; Atallah, N.; Al-Bakri, A.G.; Paradis-Bleau, C.; Zalloum, H.; Younis, K.; Levesque, R.C. *Bioorg. Med. Chem*. 2008, 16, 1218-1235.
[56] Taha, M.O.; Bustanji, Y.; Al-Bakri, A.G.; Yousef, M.; Zalloum, W.A.; Al-Masri, I.M.; Atallah, N. *J. Mol. Graphics Modell*. 2007, 25, 870–884.
[57] Al-masri, I.M.; Mohammad, M. K.; Taha, M.O. *ChemMedChem* 2008, 3, 1763–1779.
[58] Taha, M.O.; Dahabiyeh, L. A.; Bustanji, Y.; Zalloum, H.; Saleh, S. *J. Med. Chem.* 2008, 51, 6478–6494.
[59] Al-Nadaf, A.; Abu Sheikha, G.; Taha, M.O. *Bioorg. Med. Chem.* 2010, 18, 3088-115.
[60] Abu-Hammad, A. M.; Taha, M.O. *J. Chem. Inf. Model.* 2009, 49, 978–996.

[61] Abu Khalaf, R.; Abu Sheikha, G; Bustanji, Y.; Taha, M.O. *Eur. J. Med. Chem.* 2010, 45, 1598–1617.
[62] Al-Sha'er, M.; Taha, M.O. *Eur. J. Med. Chem.* 2010, 45, 4316–4330.
[63] Al-Sha'er, M.; Taha, M.O. *J. Chem. Inf. Model.* 2010, 50, 1706–1723.
[64] Taha, M.O.; Trarairah, M.; Zalloum, H.; Abu Sheikha G. *J. Mol. Graph Model.*, 2010, 28, 383-400.
[65] Abu Khalaf, R.; Abdula, A.; Mubarak, M.; Taha, M. *J. Mol. Model*. 2011, 17, 443–464.
[66] Abdula, A.; Abu Khalaf, R.; Mubarak, M.; Taha, M. *J. Comput. Chem.* 2011, 3, 463–482.
[67] CATALYST 4.11 Users' Manual (2005) Accelrys Software Inc San Diego, CA.
[68] Poptodorov K, Luu T, Langer T, Hoffmann R (2006) In: Hoffmann R D (ed) Methods and Principles in Medicinal Chemistry. Pharmacophores and Pharmacophores Searches Wiley-VCH, Weinheim, Germany
[69] Li H, Sutter J, Hoffmann R (2000) In: Güner O F (ed) Pharmacophore Perception Development and Use in Drug Design, International University Line, La Jolla, CA.
[70] Sutter J, Güner O, Hoffmann R, Li H, Waldman M (2000) In: Güner O F (ed) Pharmacophore Perception Development and Use in Drug Design, International University Line, La Jolla, CA.
[71] Discovery Studio version 25 (DS 25) User Manual (2009) Accelrys Inc, San Diego, CA
[72] Sutter J, Güner O, Hoffmann R, Li H, Waldman M (2000) In: Güner O F (ed) Pharmacophore Perception Development and Use in Drug Design, International University Line, La Jolla, CA.
[73] Kurogi Y, Güner O F (2001) Curr Med Chem 8: 1035–1055
[74] Poptodorov K, Luu T, Langer T, Hoffmann R (2006) In: Hoffmann R D (ed) Methods and Principles in Medicinal Chemistry Pharmacophores and Pharmacophores Searches, Wiley-VCH, Weinheim, Germany.
[75] Li H, Sutter J, Hoffmann R (2000) In: Güner O F (ed) Pharmacophore Perception Development and Use in Drug Design, International University Line: La Jolla, CA.
[76] Bersuker I B, Bahçeci S, Boggs JE (2000) In: Güner O F (ed) Pharmacophore Perception Development and Use in Drug Design, International University Line: La Jolla, CA.
[77] CERIUS2, QSAR Users' Manual, version 4.10; Accelrys Inc.: San Diego, CA, 2005; pp 43-88, 221-235.
[78] M.L. Verdonk, V. Berdini, M.J. Hartshorn, W.T.M. Mooij, C.W. Murray, R.D. Taylor, P. Watson *J. Chem. Inf. Model.*. 2004, 44, 793-806.
[79] J. Kirchmair, P. Markt, S. Distinto, G. Wolber, T. Langer, *J. Comput.-Aided Mol. Des.* 2008, 22, 213-228.
[80] J.J. Irwin, B.K. Shoichet, *J. Chem. Inf. Model.* 2004, 45, 177-182.
[81] N. Triballeau, F. Acher, I. Brabet, J.-P. Pin, H.-O. Bertrand, *J. Med. Chem.* 2005, 48, 2534-2547.
[82] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney. *Adv. Drug Del. Reviews*, 2001, 46, 3-26.

[83] D.F. Veber, S.R. Johnson, H.-Y. Cheng, B.R. Smith, K.W. Ward, K.D. Kopple, *J. Med. Chem.* 2002, 45, 2615-2623.



Virtual Screening

Edited by Prof. Mutasem Taha

ISBN 978-953-51-0308-0
Hard cover, 100 pages
Publisher InTech
Published online 14, March, 2012
Published in print edition March, 2012

Pharmacophore modeling, QSAR analysis, CoMFA, CoMSIA, docking and molecular dynamics simulations, are currently implemented to varying degrees in virtual screening towards discovery of new bioactive hits. Implementation of such techniques requires multidisciplinary knowledge and experience. This volume discusses established methodologies as well as new trends in virtual screening with aim of facilitating their use in drug discovery.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Mutasem O. Taha (2012). Mixing Pharmacophore Modeling and Classical QSAR Analysis as Powerful Tool for Lead Discovery, Virtual Screening, Prof. Mutasem Taha (Ed.), ISBN: 978-953-51-0308-0, InTech, Available from: http://www.intechopen.com/books/virtual-screening/mixing-pharmacophore-modeling-and-classical-qsar-analysis-as-powerful-tool-for-lead-discovery-

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821