

HQSAR Study of Tricyclic Azepine Derivatives as an EGFR (Epidermal Growth Factor Receptor) Inhibitors

Hwan Won Chung¹, Kyu Whan Lee¹,
Jung Soo Oh¹ & Seung Joo Cho¹

¹Computational Science Center, Future Fusion Technology Division, Korea Institute of Science and Technology, PO Box 131, Cheongryang, Seoul 130-650, Korea
Correspondence and requests for materials should be addressed to S. J. Cho (chosj@kist.re.kr)

Accepted 20 July 2007

Abstract

Stimulation of epidermal growth factor receptor (EGFR) is essential in signaling pathway of tumor cells. Thus, EGFR has intensely studied as an anti-cancer target. We developed hologram quantitative structure activity relationship (HQSAR) models for data set which consists of tricyclic azepine derivatives showing inhibitory activities for EGFR. The optimal HQSAR model was generated with fragment size of 6 to 7 while differentiating fragments having different atom and connectivity. The model showed cross-validated q^2 value of 0.61 and non-cross-validated r^2 value of 0.93. When the model was validated with an external set excluding one outlier, it gave predictive r^2 value of 0.43. The contribution maps generated from this model were used to interpret the atomic contribution of each atom to the overall inhibition activity. This can be used to find more efficient EGFR inhibitors.

Keywords: HQSAR, EGFR, Kinase inhibitor, Tricyclic azepine

Anticancer drugs have been conventionally developed for the control of DNA synthesis and function or cell mitosis¹. Such cytotoxic drugs may have drawbacks in their efficacy of cell death and in its selectivity between malignant tumor cells and normal cells¹. A signaling pathway consists of three steps: binding of a small ligand to an extracellular receptor, transduction of an external signal into a cell, and alteration of protein-protein interactions. This signaling flow regulates all of cell behaviors like proliferation, growth, and differentiation. If the signaling pathways fail to control normal cell proliferation and

survival, many cancers can appear due to that disruption². Usually, the over-expression of special growth-factor-receptor tyrosine kinases and their mutated forms may result in the unrestrained activation of that pathway³. Thus, the use of signaling pathways as molecular targets can be a new gate to anticancer drug development¹.

Protein-tyrosine kinases (PTKs) have become important targets for anticancer drugs because of their role as regulators of intracellular signal-transduction pathways⁴. EGFR (epidermal growth factor receptor), a kind of the PTK family, is one of the most intensively studied targets for anticancer drugs. Stimulation of the EGFR signaling pathway in malignant tumor cells can produce increased cell proliferation, angiogenesis, and metastasis, and decreased apoptosis⁶. Once binding of an epidermal growth factor (EGF) initiates the activation of the EGFR, then the receptor can develop into a homodimer with another EGFR monomer or a heterodimer with different member of the erbB family, which consists of four similar receptors. After dimerization, the intrinsic kinase activity is increased and tyrosine autophosphorylation takes place⁶. EGF receptors are exceptional in that their stimulation of kinase activity is not from the activation loop autophosphorylation, but from the generation of a cytoplasmic domain dimer. So, autophosphorylated but monomeric EGF receptors are not activated. Most of the tyrosine sites in the EGF receptor exist in the carboxy-terminal region of the receptor⁵. Furthermore, the activation loop of the EGFR not having a phosphate group takes a structure analogous to that of the kinase domain from the insulin receptor which is autophosphorylated⁷.

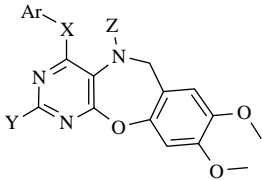
EGFR-directed therapies have distinct effects on the treatment of tumor cells so that various areas of the EGFR have been used for the development of inhibitors: "the extracellular ligand-binding domain, the intracellular tyrosine kinase domain, the ligand, or the synthesis of the EGFR from DNA"⁶. According to various target areas, there exist several EGFR-targeted strategies: monoclonal antibodies, bispecific antibodies, EGFR-TKIs (tyrosine kinase inhibitors), recombinant vaccine, and antisense oligonucleotides⁶. The ATP-binding pocket of tyrosine kinase domain within EGFR can be a good target area for the design of inhibitors suppressing the unregulated activation of EGFR. A strategy using such a pocket with a small

molecule is promising⁸.

Hologram quantitative structure-activity relationship (HQSAR) is a relatively new method for structure-activity relationships which utilizes weighted 2D fingerprints in conjunction with the PLS statistics. HQSAR generates a predictive model of biological activity with its predictive variables based on distinctive fragment fingerprints (molecular holograms)⁹. Generally, molecular function is resulted from its structure; therefore, 2D fingerprint, converted representation of a molecular substructure, can have prediction power for a molecule¹⁰. HQSAR model can predict biological activities of new compounds, and the model also gives some indications about which substituents or structures should be modified to increase the activity of compounds⁹. Since the HQSAR is so fast in model generation and do not need molecular alignment, the method can be used for both small and large data sets, and HQSAR can also support database searching⁹.

Smith *et al.* reported that oxazepine derivatives inhibit EGFR tyrosine kinase⁸. In this study, HQSAR⁹ technique has been applied to investigate the factors affecting inhibitory activity of tricyclic azepine derivatives.

Table 1. Oxazepines as inhibitors of EGFR and their inhibition activities⁸.

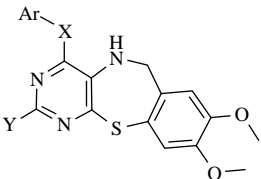


Compound	Ar	X	Y	Z	pIC ₅₀
1c	3-Br-Ph	NH	H	H	6.52
2a	3-Me-Ph	NH	H	H	5.13
2b	3-Ethynyl-Ph	NH	H	H	5.49
2c	4-Br-Ph	NH	H	H	5.06
2d	4-F-Ph	NH	H	H	5.15
2e	3-Cl-4-F-Ph	NH	H	H	5.92
2f*	3-Cl-2-F-Ph	NH	H	H	6.52
2g	5-Cl-2-F-Ph	NH	H	H	5.96
2h*	2-Cl-4-F-Ph	NH	H	H	5.47
2i*	6-Indazolyl	NH	H	H	5.92
2j	2-Naphthyl	NH	H	H	6.3
2k	6-Benzthiazolyl	NH	H	H	5.25
2n	3-Br-Ph	O	H	H	6.15
2o	3-Cl-4-F-Ph	O	H	H	5.92
2p	3-Br-Ph	O	H	Me	5.34
2q	3-Cl-4-F-Ph	O	H	Me	4.97
2r	3-Br-Ph	S	H	H	5.89
2s*	3-Cl-Ph	S	H	H	

HQSAR Model

To find good model from HQSAR analysis, we examined the influence of the fragment distinction and the fragment size on the important statistical parameters. HQSAR investigation was carried out using the following distinctions: atoms (A), bonds (B), connections (Co), hydrogen atoms (H) and donor and acceptor (DA). With the default fragment size (4-7), some combinations of this distinction information were taken into account as follows: A, A/B, A/B/Co, A/B/Co/H, A/Co, Co, B/Co, and A/Co/DA. After HQSAR analysis was accomplished using the 12 default hologram lengths of 53, 59, 61, 71, 83, 97, 151, 199, 257, 307, 353, and 401 bins for each parameter set, the best models and its optimal number of components (LV) were chosen based on the least cross validated standard error SE_{cv}. The results of HQSAR analyses for the training data set using some fragment distinction information are shown in Table 3. The best model was produced from fragment distinction of atoms and connections, and the model showed cross-validated r² (q²) value of 0.572 and non-cross-validated r² value of 0.928. When only A or A/B parameters were used as fragment distinction, q² values were 0.471 and 0.408, respectively; on the other hand, when only connectivity used as fragment distinction, q² value was 0.572. This means that

Table 2. Thiazepines as inhibitors of EGFR and their inhibition activities⁸.



Compound	Ar	X	Y	pIC ₅₀
3a	3-Br-Ph	NH	H	5.23
3b*	3-Cl-4-F-Ph	NH	H	5.09
3c	3-Cl-2-F-Ph	NH	H	4.2
3d*	6-Indazolyl	NH	H	4.85
3g	3-Cl-Ph	NMe	H	4.57
3h*	3-Br-Ph	NH	Me	4.51
3j	3-Br-4-Me-Ph	NH	Me	4.34
3m	2-Naphthyl	NH	Me	4.77
3n	5-Benzimidazolyl	NH	Me	4.84
3o*	6-Benzthiazolyl	NH	Me	4.82
3p	6-Indazolyl	NH	Me	4.61
3q	5-Indazolyl	NH	Me	4.79
3r*	3-Br-Ph	O	H	4.66
3s	3-Cl-2-F-Ph	O	H	4.71
3t	3-Br-Ph	S	H	5.8
3u*	3-Cl-Ph	S	H	5.54

Table 3. HQSAR analyses for various fragment distinctions using default fragment size (4-7); (LVmax=8), model validation by LOO (Leave-One-Out).

Distinction	Fragment					
	q ²	SE _{cv}	r ²	SEE	LV	Length
A	0.471	0.515	0.738	0.362	4	97
A/B	0.408	0.544	0.741	0.36	4	199
A/B/Co	0.565	0.466	0.942	0.17	4	83
A/B/Co/H	0.558	0.497	0.956	0.157	6	59
A/Co	0.576	0.461	0.928	0.189	4	71
Co	0.572	0.476	0.937	0.183	5	151
B/Co	0.468	0.503	0.889	0.23	3	199
A/Co/DA	0.383	0.529	0.508	0.472	2	71

Table 4. HQSAR analysis for the influence of various fragment sizes using the best fragment distinction (A/Co), LVmax (4).

Size	Fragment					
	q ²	SE _{cv}	r ²	SEE	LV	Length
1-2	0.306	0.548	0.427	0.498	1	97
1-3	0.470	0.502	0.781	0.323	3	53
1-10	0.435	0.494	0.554	0.439	1	83
2-10	0.436	0.494	0.555	0.439	1	83
3-9	0.432	0.496	0.544	0.444	1	83
3-10	0.438	0.556	0.556	0.438	1	83
4-7	0.576	0.461	0.928	0.189	4	71
5-7	0.585	0.456	0.928	0.191	4	71
5-10	0.436	0.494	0.554	0.439	1	83
6-7	0.612	0.441	0.931	0.185	4	71
7-9	0.438	0.493	0.547	0.443	1	71
8-9	0.430	0.497	0.548	0.442	1	53
9-10	0.446	0.490	0.562	0.435	1	97s

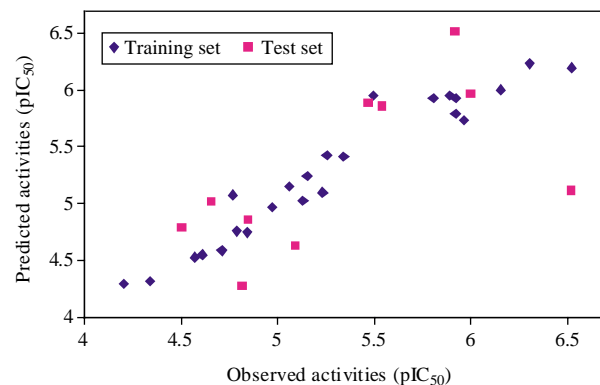
connectivity was more important than atoms or bonds in this analysis, and when both atoms and connectivity (A/Co) were used, the best model was introduced.

We attempted to find a model with better statistics by changing fragment sizes, and analyses have been repeated with different fragment sizes using the previous best model's fragment distinction parameter (A/Co). The evaluated fragment sizes and their statistical results are shown in Table 4. Although hologram length of the best model was not changed, two fragment sizes (5-7, 6-7) showing better q² values than the previous best model were generated, and the model with its q² value (0.612) and r² value (0.931) was the best model found.

External test set (marked asterisk in Table 1, 2) was used to make sure the prediction power of the best model made from the 24 training set molecules. The test set's pIC₅₀ values were calculated, and the predictive r² value was 0.43 when one outlier data of

Table 5. Observed versus predicted activities (pIC₅₀) with residuals by HQSAR.

Compound	Observed	Predicted	Residual
Training			
1c	6.52	6.2	0.32
2a	5.13	5.03	0.1
2b	5.49	5.95	-0.46
2c	5.06	5.15	-0.09
2d	5.15	5.25	-0.1
2e	5.92	5.93	-0.01
2g	5.96	5.73	0.23
2j	6.3	6.24	0.06
2k	5.25	5.43	-0.18
2n	6.15	6	0.15
2o	5.92	5.79	0.13
2p	5.34	5.42	-0.08
2q	4.97	4.97	0
2r	5.89	5.95	-0.06
3a	5.23	5.1	0.13
3c	4.2	4.29	-0.09
3g	4.57	4.53	0.04
3m	4.34	4.32	0.02
3n	4.77	5.08	-0.31
3p	4.84	4.74	0.1
3q	4.61	4.55	0.06
3s	4.79	4.76	0.03
3t	4.71	4.59	0.12
Test			
2f	6.52	5.117	1.403
2h	5.47	5.879	-0.409
2i	5.92	6.519	-0.599
2s	6	5.963	0.037
3b	5.09	4.628	0.462
3d	4.85	4.857	-0.007
3h	4.51	4.786	-0.276
3o	4.82	4.268	0.552
3r	4.66	5.008	-0.348
3u	5.54	5.855	-0.315


Figure 1. Plot of observed versus predictive activities (pIC₅₀).

2f compound was excluded. The observed and predictive activities of both training set and test set were

shown in Table 5. Figure 1 shows the plot diagram of observed versus predicted activities of both training set and test set.

Atomic Contribution

In HQSAR, contribution map represents one model graphically and the color of each atom in the map reflects the contribution of that atom to the compound's inhibitory activity. Colors at the red end of the spectrum (red, red orange, and orange) reflect poor (or negative) contributions, while colors at the green end (yellow, green blue, and blue) reflect favorable (positive) contributions. White colors reflect intermediate contributions.

Figure 2 shows the atomic contribution of the model compound (1c) to its inhibition activity. It indicates that pyrimidine ring mainly contributes the overall activity of the compound. This corresponds to the bad effect resulted from the methyl substitution of pyrimidine ring. When para position of aniline ring was substituted to halogen group, it did not give any contribution to the bioactivity (refer to Figures 2, 3).

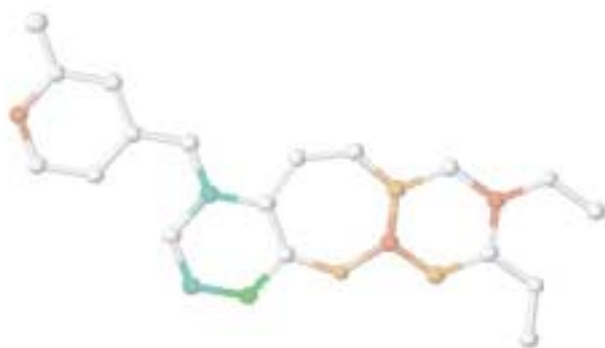


Figure 2. Contribution map of compound 1c.

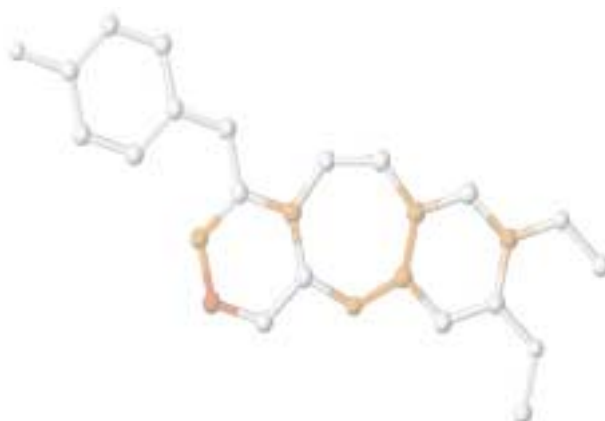


Figure 3. Contribution map of compound 2d.

NH group of the 7-membered ring (of the oxazepine and thiazepine) is also important for good activity because its substitution to methyl showed significant

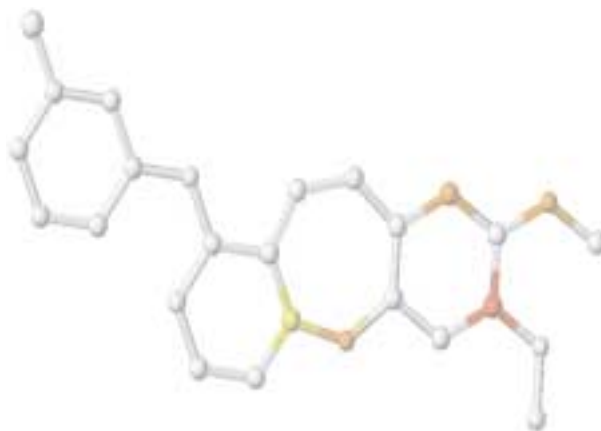


Figure 4. Contribution map of 2n.

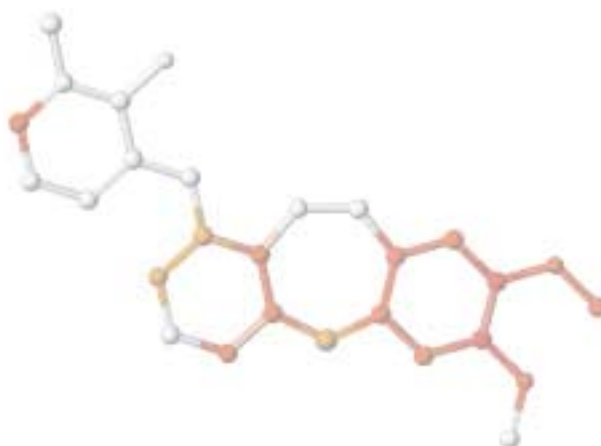


Figure 5. Contribution map of compound 3c.



Figure 6. Contribution map of compound 3p.

decrease of the model compound (Figure 6). Figure 4 indicates that NH-group of aniline ring can be changed to oxygen without significant decrease of the inhibition activity. Generally, thiazepine ring represents considerable decrease in activity than oxazepine ring because sulfur atom in thiazepine might have some steric hindrance to binding of model compound (Refer to Figures 5, 6).

Discussion

HQSAR analysis was performed to investigate the 2D QSAR of tricyclic azepine derivatives as EGFR inhibitors. When atoms and connectivity information were used, the best model was introduced, and it was further optimized with their fragment sizes changed. The optimal model showed q^2 value of 0.612 and r^2 value of 0.928. Due to structural variety of the data set used in this study, this model did not show high prediction ability for the external data set, however, the model was statistically robust with a good explanatory power. The contribution maps generated from this model showed that pyrimidine ring and NH-group of azepine ring are critical for good inhibition activity. Generally, oxazepine ring was better than thiazepine ring in its biological activity. This analysis will be useful to find promising analogues for EGFR inhibitors.

Methods

Data Sets

Thirty-four tricyclic azepine derivatives, a subset of the dataset published by Smith *et al.* with two different scaffolds⁸, were used for the HQSAR analysis. Table 1 and 2 summarizes the structures of the molecules and their inhibitory activities (pIC_{50} values, μM). In this dataset, twenty-four compounds were selected as a training set to generate QSAR models and the other ten compounds (Table 1-2 marked with asterisk) were used as a test set for model validation. The inhibitory activity values were transformed from IC_{50} values to pIC_{50} ($-\log \text{IC}_{50}$) values by scaling, then they were evaluated as statistical variables in the QSAR investigations.

HQSAR Analysis

HQSAR analysis requires input molecular structures and their biological activities, and it proceeds to the generation of molecular holograms and model generation. HQSAR analysis consists of three main activities: the generation of structural fragments for

all molecules in the training set, derivation of the molecular holograms by hashing to defined hologram lengths, and generation of HQSAR models following its validation¹⁰⁻¹².

The input molecules were separated into all possible structural fragments including branched, cyclic, and overlapping fragments and their atom sizes range from minimum (M) to maximum (N). Molecular hologram was derived from the modification of bit string and it consisted of an array of integers indicating how many times each fragment was put into each bin. Since all molecular structures should be encoded into molecular holograms, all structural fragments were expressed as SLN strings¹¹, and they were transformed into pseudo-random integers by CRC (Cyclic Redundancy Check) algorithm. These numbers were hashed into hologram array of the specified length (L) in the range 1 to L. With all generated fragments hashed into hologram array, this hologram array contains bin occupancies which are the descriptor variables. The use of hashing reduces the size of molecular hologram and decreases the computation time of HQSAR, and it is necessary because in some cases there are more unique fragments than fingerprint positions (bins). However, hashing also makes 'fragment collision' problem, so different unique fragments can hash into the same bin, which is related to chance correlation. To reduce the fragment collision, hologram lengths (the value of L) were selected to be prime numbers (default values of which are 53, 59, 61, 71, 83, 97, 151, 199, 257, 307, 353, and 401). As one hologram was created for one molecule, then all the other holograms corresponding to remaining compounds were generated by repeating previous process. In the HQSAR procedure, fragment sizes (M, N) and hologram length (L) mainly control the creation of molecular hologram, so many combinations of fragment sizes and hologram lengths were used to find the set of parameters (M, N, and L) which resulted in the best HQSAR. HQSAR models were generated from biological activities and molecular holograms in conjunction with the PLS (Partial Least Square) method. Cross-validation with leave one out method was also performed to determine the number of components which gave the best refined model. From the stored PLS results, best hologram length and their fragment sizes were chosen and reported. Selected optimal model produced a statistical equation relating the molecular hologram bin values to the biological activity of each compound in the data set.

$$\text{Activity}_i = c_0 + \sum_L c_{il} x_{il}$$

Where x_{il} is the occupancy value of the molecular

hologram of compound i at position l , c_{il} is the coefficient for that bin (position) derived from the PLS analysis, L is the length of the hologram, Activity is the biological activity, and c_0 is a constant. PLS solution coefficients from the optimally selected structure-activity relationship were used to predict the activities of test compounds and to visualize relative contribution of each molecule of data set.

HQSAR Contribution Maps

In contribution map, the result of HQSAR analysis was shown as a color-coded structure diagram in which each atom had the color corresponding to the intensity of its contribution to the molecule's overall activity¹². If one atom had weak contribution, it was displayed by the colors of red, red-orange and orange which were located at the red end of color spectrum. On the contrary, an atom with strong contribution was drawn by the colors of yellow, green-blue and green which were located at the green end of the color spectrum. Intermediate contributions were depicted in white color.

The atomic contribution for each atom needs calculation using PLS coefficients and the hologram of optimal model. When several fragments are positioned in a fingerprint bin by hashing, then every atom of them will take the same part in the PLS coefficient because there does not exist the way to discriminate its contribution to the statistic. A weighting value for each atom equals to the PLS coefficient divided by the number of atoms in the fragments pertaining to the location. In each position in the fingerprint, all atoms get their weighting value from such a procedure, so individual atoms will have received weights from many different fingerprint positions. Each atom get its contribution color scaled from the minimum and maximum weighting values of whole molecules of the data set¹².

Predictive r Squared (r^2 pred)

To validate the derived HQSAR models, biological activities of an external test set were predicted using models derived from the training set. By predicting the biological activities of an external test set, we tried to check the validity of the selected HQSAR model which is derived from the training set. Predictive r^2 value was used as an expression of the prediction power of the model and it is similar to

cross-validated r^2 (q^2). This predictive r^2 was calculated by next formula¹⁰

$$r^2_{\text{pred}} = (\text{SD} - \text{PRESS}) / \text{SD}$$

where SD was the sum of squared deviation between the biological activities of the test set molecule and the mean activities of the training set molecules and PRESS was the sum of squared deviations between the observed and the predicted activities of the test molecules.

References

1. Kurup, A., Garg, R. & Hansch, C. Comparative QSAR study of tyrosine kinase inhibitors. *Chem Rev* **101**:2573-2600 (2001).
2. Cooper, G. M. The cell-a molecular approach, in, Edn. Second (Sinauer Associates, 2000).
3. Cohen, P. Protein kinases-the major drug targets of the twenty-first century? *Nature Reviews Drug Discovery* **1**:309-315 (2002).
4. Blume-Jensen, P. & Hunter, T. Oncogenic kinase signalling. *Nature* **411**:355-365 (2001).
5. Hubbard, S. R. & Till, J. H. Protein tyrosine kinase structure and function. *Annual Review of Biochemistry* **69**:373-398 (2000).
6. Baselga, J. Why the epidermal growth factor receptor? The Rationale for Cancer Therapy. *Oncologist* **7**:2-8 (2002).
7. Stamos, J., Sliwkowski, M. X. & Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J Biol Chem* **277**:46265-46272 (2002).
8. Smith Ii, L. *et al.* Tricyclic azepine derivatives: Pyrimido[4, 5-b]-1, 4-benzoxazepines as a novel class of epidermal growth factor receptor kinase inhibitors. *Bioorganic & Medicinal Chemistry Letters* **16**:1643-1646 (2006).
9. David, R. Lowis. HQSAR: A new, highly predictive QSAR technique. Tripos Technical Notes, Tripos., (1997).
10. Tripos Bookshelf 7.3, Tripos Inc., (1699).
11. Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J Chem Inf Model* **38**:379-386 (1998).
12. John Robert, H. & Trevor William, H. Tripos, Inc, St. Louis, MO (US). Molecular Hologram QSAR. United States patent US 6208942 B1. (Feb. 10. 1998)