# *Identification of Potent EGFR Inhibitors from TCM Database@Taiwan*

# Identification of Potent EGFR Inhibitors from TCM Database@Taiwan

Shun-Chieh Yang[1], Su-Sen Chang[1], Hsin-Yi Chen[2], Calvin Yu-Chian Chen[1,2,3]*

1 Laboratory of Computational and Systems Biology, School of Chinese Medicine, China Medical University, Taichung, Taiwan, 2 Department of Bioinformatics, Asia University, Taichung, Taiwan, 3 China Medical University Beigang Hospital, Yunlin, Taiwan

## Abstract

Overexpression of epidermal growth factor receptor (EGFR) has been associated with cancer. Targeted inhibition of the EGFR pathway has been shown to limit proliferation of cancerous cells. Hence, we employed Traditional Chinese Medicine Database (TCM Database@Taiwan) (http://tcm.cmu.edu.tw) to identify potential EGFR inhibitor. Multiple Linear Regression (MLR), Support Vector Machine (SVM), Comparative Molecular Field Analysis (CoMFA), and Comparative Molecular Similarities Indices Analysis (CoMSIA) models were generated using a training set of EGFR ligands of known inhibitory activities. The top four TCM candidates based on DockScore were 2-O-caffeoyl tartaric acid, Emitine, Rosmaricine, and 2-O-feruloyl tartaric acid, and all had higher binding affinities than the control Iressa®. The TCM candidates had interactions with Asp855, Lys716, and Lys728, all which are residues of the protein kinase binding site. Validated MLR ($r^2 = 0.7858$) and SVM ($r^2 = 0.8754$) models predicted good bioactivity for the TCM candidates. In addition, the TCM candidates contoured well to the 3D-Quantitative Structure-Activity Relationship (3D-QSAR) map derived from the CoMFA ($q^2 = 0.721$, $r^2 = 0.986$) and CoMSIA ($q^2 = 0.662$, $r^2 = 0.988$) models. The steric field, hydrophobic field, and H-bond of the 3D-QSAR map were well matched by each TCM candidate. Molecular docking indicated that all TCM candidates formed H-bonds within the EGFR protein kinase domain. Based on the different structures, H-bonds were formed at either Asp855 or Lys716/Lys728. The compounds remained stable throughout molecular dynamics (MD) simulation. Based on the results of this study, 2-O-caffeoyl tartaric acid, Emitine, Rosmaricine, and 2-O-feruloyl tartaric acid are suggested to be potential EGFR inhibitors.

**Citation:** Yang S-C, Chang S-S, Chen H-Y, Chen CY-C (2011) Identification of Potent EGFR Inhibitors from TCM Database@Taiwan. PLoS Comput Biol 7(10): e1002189. doi:10.1371/journal.pcbi.1002189

**Editor:** James M. Briggs, University of Houston, United States of America

**Received** June 14, 2011; **Accepted** July 25, 2011; **Published** October 13, 2011

**Funding:** The research was supported by grants from the National Science Council of Taiwan (NSC 99-2221-E-039-013-), Committee on Chinese Medicine and Pharmacy (CCMP100-RD-030), China Medical University and Asia University (CMU98-TCM, CMU99-TCM, CMU99-S-02, CMU99-ASIA-25, CMU99-ASIA-26 CMU99-ASIA-27 CMU99-ASIA-28). This study is also supported in part by Taiwan Department of Health Clinical Trial and Research Center of Excellence (DOH100-TD-B-111-004) and Taiwan Department of Health Cancer Research Center of Excellence (DOH100-TD-C-111-005). We are grateful to the Asia University cloud-computing facilities. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: ycc929@MIT.EDU; ycc@mail.cmu.edu.tw

## Introduction

Target-specific therapies have generated much attention in addition to conventional cancer treatments [1–3]. By targeting key molecules essential for cellular function, replication, or tumorigenesis, such therapies may exert cytostatic or cytotoxic effects on tumors while minimizing nonspecific toxicities associated with chemotherapy or irradiation [4].

The epidermal growth factor receptor (EGFR) signaling pathway is one of the most important pathways in mammalian cells [5]. Specific ligands, such as epidermal growth factor (EGF) and transforming growth factor alpha (TGFα), bind and activate EGFR, triggering autophosphorylation of the intracytoplasmic EGFR tyrosine kinase domain [6,7]. The phosphorylated tyrosine kinase residues serve as binding sites for signal transducers and activators of intracellular substrates, which then stimulate intracellular signal transduction cascades that upregulate biological processes such as gene expression, proliferation, angiogenesis, and inhibition of apoptosis [8]. EGFR overexpression has been shown to activate downstream signaling pathways, resulting in cells that have aggressive growth and invasive characteristics [9]. Tumor cell motility, adhesion, metastasis, and angiogenesis have also been associated with stimulated EGFR pathways [10–12]. Since EGFR over-expression often differentiates tumor cells from normal cells, it is possible for EGFR inhibitory molecules to act on tumor cells and attenuate their proliferation rates [4].

Several tyrosine kinase inhibitors were approved for clinical use. Iressa® (gefitinib) is highly selective for EGFR tyrosine kinase and is commonly used for treating lung cancer [13]. EGFR downstream signaling is competitively inhibited by Iressa® at its ATP binding site [14]. Other therapeutic agents with inhibitory mechanisms similar to Iressa® include Erlotinib (Tarceva®) against non-small cell lung cancer (NSCLC) and pancreatic cancer [15,16], and Vandetanib (Zactima®) against late stage medullary thyroid cancer [17]. Lapatinib (Tykerb®) is a dual inhibitor of EGFR and HER2 tyrosine kinases approved for metastatic breast cancer [18,19]. Though the effect of Iressa® on lung cancer has been well established, severe side effects has also been reported [20]. Adverse reactions listed under Iressa® product information include diarrhea, skin rash and dryness, nausea, vomiting, haemorrhage, anorexia, asthenia, and in some cases, interstitial lung disease with fatal outcomes [21]. The adverse effects of

### Author Summary

Tumor growth is associated with overexpression of epidermal growth factors receptors. Targeted control of EGFR by EGFR inhibitors is an attractive therapy alternative to conventional cancer treatment that offers specificity and reduced adverse effects. The purpose of this study was to identify natural compounds from traditional Chinese medicine that may be used as EGFR inhibitors. The top four TCM compounds with the highest binding affinity to EGFR were selected and their suitability as EGFR inhibitors confirmed with different statistical prediction models. The candidate compounds had higher bioactivity than Iressa®, the drug that is clinically used. The TCM compounds also met key structural components that were characteristic among known inhibitors. In addition, the binding between TCM compounds and EGFR were stable which is a fundamental requirement for any targeting drug. Results from bioactivity prediction, structural component matching, and binding stability all point to the possibility of these TCM compounds as suitable EGFR inhibitor candidates.

available treatments necessitate continuous search efforts for alternatives with less toxicity.

Computational predictions in biology and biomedicine are of significant importance for generating useful data which otherwise be time-consuming and costly through experiments alone [3,22–27]. Computational predictions, combined with information derived from structural bioinformatics analysis, can provide useful insights and timely information for both basic research and drug development [28,29]. Much cutting-edge cancer drug development has been conducted through the use of computational bioinformatics and modeling [30–37]. The powerful ability of modern computational prediction and bioinformatics were adopted in this research to search for novel EGFR inhibitors.

Traditional Chinese medicines (TCM) are natural substances with therapeutic effects on many diseases [38–40]. The vast number of TCM represents a rich resource that can be explored for drug development. We had investigated kinase inhibitor candidates from TCM targeting HER2 and HSP90 receptors before [28,41–42]. Though EGFR kinase inhibitors have been investigated through different screening and modeling scenarios [43–47], none from TCM compounds has been reported to date. This study utilized the world's largest TCM Database@Taiwan [48] to screen for potential EGFR inhibitors from TCM compounds and applied structure- and ligand-based methods to evaluate the suitability of candidates as EGFR inhibitors.

## Materials and Methods

A useful predictor for a biological system should include the following steps [49]: (i) selection of a valid dataset to train and test the predictor; (ii) formulate samples with an effective mathematical expression that reflects intrinsic correlation with the attribute to be predicted; (iii) develop a powerful algorithm to operate the prediction; (iv) objectively evaluate accuracy of the predictor through cross-validation tests. The experimental design of the current study follows these guidelines and details are explained in the following sections.

### EGFR Protein Sequence, Structure, and Characteristics

The EGFR protein sequence (EGFR_HUMAN, P00533) used in this study was obtained from Swiss-Prot [50], and the 3D structure (PDB: 2ITY) [51] used for analyses was downloaded from Protein Data Bank. The tyrosine kinase was encoded from Phe712-Leu979, and the ATP binding site was located at Leu718–Val726.

### Candidate Screening and Docking Studies

The Traditional Chinese Medicine (TCM Database@Taiwan, url: ) database (http://tcm.cmu.edu.tw) was used to screen for potential EGFR inhibitors from more than 20,000 compounds within the database. All compounds were operated using the Prepare Ligands module with Lipinski's rule of five using Discovery Studio 2.5 (DS 2.5; Accelrys Inc., San Diego, CA). Iressa® was selected as the control. The LigandFit program (DS 2.5) was used to locate the best docking pose for different confirmations under the Chemistry at HARvard Macromolecular Mechanics (CHARMm) force field [52]. Results for the docking studies were ranked according to Dock Score.

### Descriptor Generation Using Genetic Function Approximation (GFA) Algorithm

Twenty ligands with demonstrated inhibition against EGFR were used in this study (Table S1) [53]. Descriptors for each ligand were identified using the Calculate Molecular Properties program in DS 2.5. Predictive models containing five optimum descriptors were generated using the Genetic Function Approximation (GFA) algorithm. Descriptors in the model with the highest $r^2$ value were used to construct ligand activity prediction models.

### Ligand Activity Predictions Using Multiple Linear Regression (MLR) and Support Vector Machines (SVM)

A MLR model using the descriptors from the top GFA algorithm was constructed using Matlab Statistics Toolbox (MathWorks, Natick, MA) and validated using MLR Leave-One-Out validation [54]. The MLR model was trained with 17 randomly selected ligands with EGFR inhibitory activity (Table S1) and used to predict the activity ($pIC_{50}$) of the control and TCM candidates.

The identical descriptors were normalized to the range of $[-1,+1]$ and plugged into the libSVM program to generate a SVM prediction model[55]. Following model training with the 17-ligand training set, the SVM model was used to predict the activity of the control and TCM candidates.

### 3D Quantitative Structure-Activity Relationship (QSAR) Model

Ligands used in the previous sections were also used for 3D-QSAR analysis. The 2-dimensional (2D) and 3-dimensional (3D) ligand structures were drawn with ChemBioOffice 2008 (PerkinElmer Inc., Cambridge, MA) under a Molecular Mechanics 2 (MM2) force field. Following ligand alignment, Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarities Indices Analysis (CoMSIA) models were constructed using partial least squares statistical method (PLS). Cross-Validated (CV) correlation coefficient ($q^2$) and non-cross-validation correlation coefficient ($r^2$) were used to validate the models. Biological activities of Iressa® and TCM candidate compounds were predicted using the generated 3D-QSAR contour map.

### Molecular Dynamics Simulation

Molecular dynamics (MD) of Iressa® and the TCM candidates were simulated using DS2.5 Standard Dynamics Cascade and Dynamics package. Sample preparation was conducted under the following parameters: [minimization] steepest descent and conjugate gradient: each with maximum steps of 500, [heating time] 50 ps, [equilibration time] 200 ps. The simulations were produced with a total production time of 20 ns with NVT, constant temperature dynamics of Berendsen weak coupling method, a temperature decay

**Table 1.** Docking score and biological activity predictions of top TCM candidates in comparison with the control.

| | | Predicted Activity ($pIC_{50}$) | |
|---|---|---|---|
| **Compounds** | **Dock Score** | **MLR** | **SVM** |
| 2-O-Caffeoyl tartaric acid | 118.065 | 8.386 | 7.041 |
| Emetine | 113.424 | 5.548 | 7.140 |
| Rosmaricine | 105.149 | 4.543 | 7.463 |
| 2-O-Feruloyl tartaric acid | 104.284 | 8.359 | 7.242 |
| Iressa* | 60.791 | 6.715 | 5.110 |

*control.

doi:10.1371/journal.pcbi.1002189.t001

time of 0.4 ps, and a target temperature of 310K. Root mean square deviations (RMSD) of protein-ligand complex and individual ligands, total energy of protein-ligand complex, hydrogen bond (H-bond), and H-bond distance were analyzed using the Analyze Trajectory function following MD simulation.

## Results/Discussion

### Candidate Screening and Docking Studies

The top four TCM candidates with the highest Dock Score were 2-O-caffeoyl tartaric acid, Emitine, Rosmaricine, and 2-O-feruloyl tartaric acid (Table 1). Corresponding scaffolds of the top TCM candidates are illustrated in Figure 1. Iressa®, Emitine, and Rosmaricine had amine groups available for H bonding whereas 2-O-Caffeoyl tartaric acid and 2-O-feruloyl tartaric acid had carbonyl groups. The different residues available for H bonding resulted in different binding poses (Figure 2). Binding of Iressa® (Figure 2a), Emitine (Figure 2c), and Rosmaricine (Figure 2e) to tyrosine kinase were located within the pocket, with H-bonds formed between the amine group of the ligand compounds and the carboxyl group of Asp855. 2-O-Caffeoyl tartaric acid (Figure 2b) and 2-O-feruloyl tartaric acid (Figure 2e) docked outside the tyrosine kinase pocket and formed multiple H-bonds through their carboxyl groups with Lys716 and Lys728. The binding location of 2-O-caffeoyl tartaric acid and 2-O-feruloyl tartaric acid directly blocks the ATP binding site of tyrosine kinase located from Leu718–Val726. Dock scores of each TCM candidate is given in Table 1. All candidates have higher dock scores than Iressa®, indicating higher binding affinities to the tyrosine kinase receptor than Iressa®.

### Ligand Activity Predictions Using MLR and SVM

Representative descriptors from the top GFA algorithm include: Num_H_Acceptors_Lipinski (equivalent of N+O count), Molecular_SurfaceArea (the total surface area for each molecule using a 2D approximation), Kappa_1 (Kappa Shape Indices), PMI_Y (principle moment of inertia Y-component), and Shadow Xlength (length of molecule in the X dimension). The descriptors were validated using Leave-One-Out method which is the most objective of all available cross-validation methods [56]. The MLR model established with the determined descriptors was:

$$
\begin{aligned}
pIC_{50} = {} & 6.7148 - 0.4477 * \mathrm{Num\_H\_Acceptors\_Lipinski} - 0.0800 * \\
& \mathrm{Molecular\_SurfaceArea} + 1.5172 * \mathrm{Kappa\_1} + 0.00027466 * \\
& \mathrm{PMI\_Y} - 0.0452 * \mathrm{Shadow\_Xlength}
\end{aligned}
$$

The SVM model was also established with the five identified descriptors using libSVM.

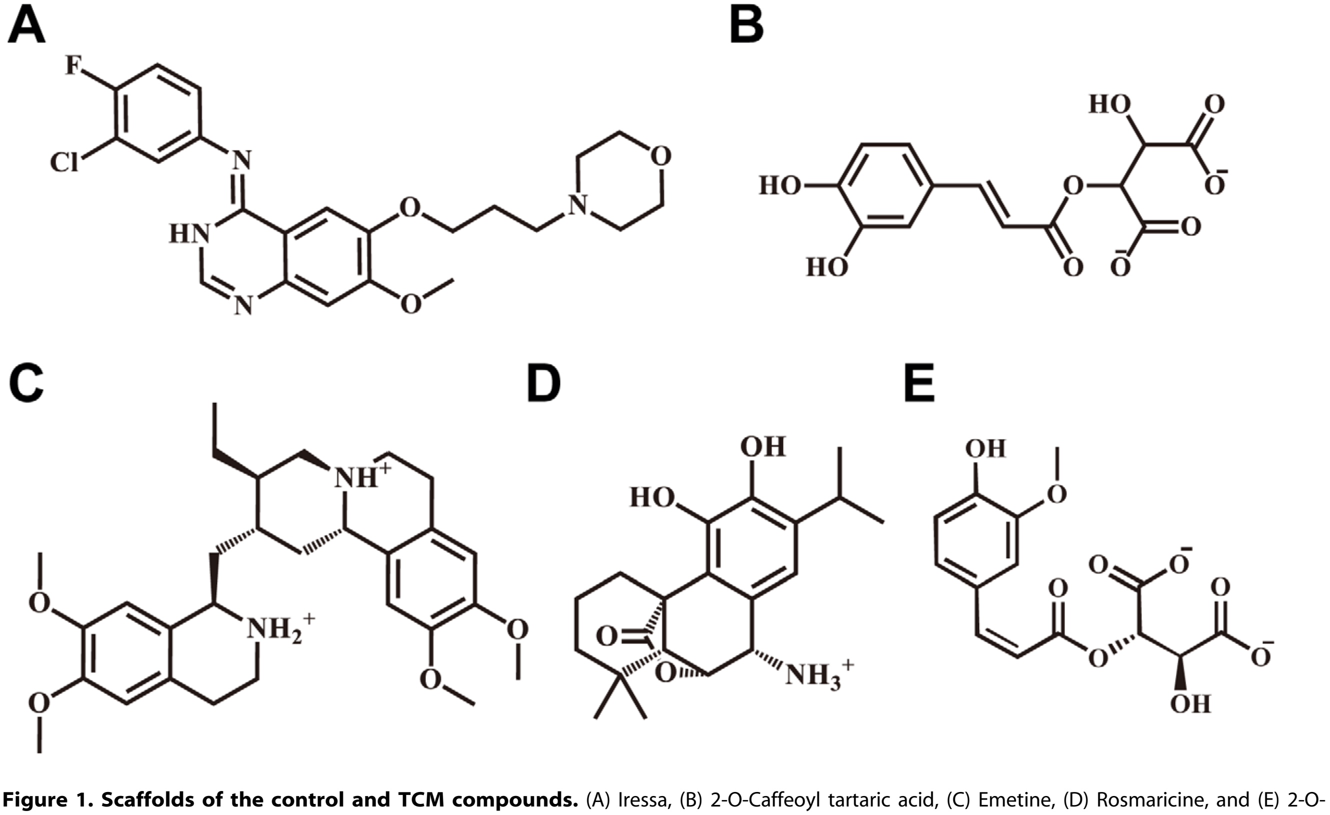

**Figure 1. Scaffolds of the control and TCM compounds.** (A) Iressa, (B) 2-O-Caffeoyl tartaric acid, (C) Emetine, (D) Rosmaricine, and (E) 2-O-Feruloyl tartaric acid.

doi:10.1371/journal.pcbi.1002189.g001

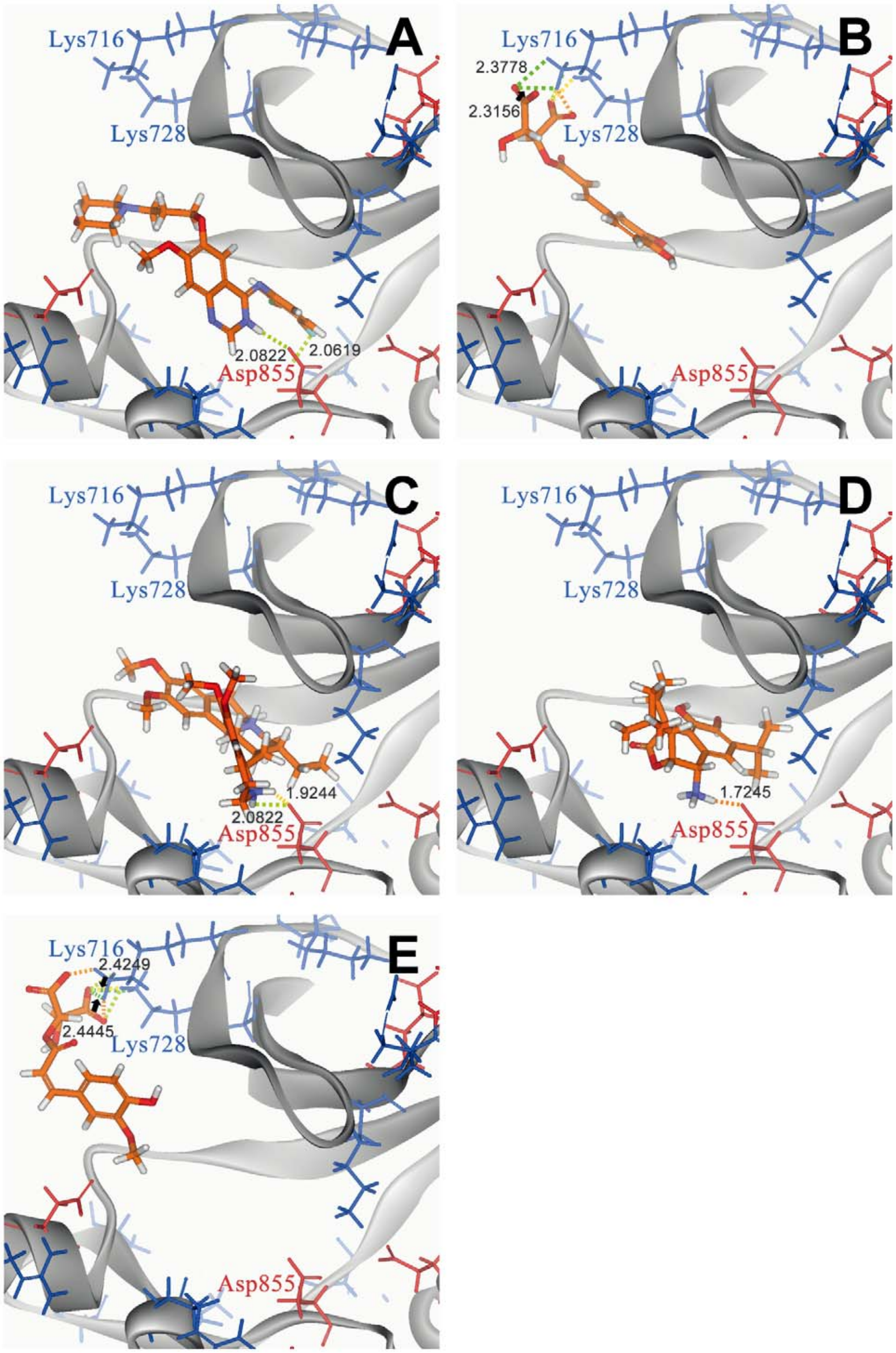

**Figure 2. Docking pose of different compounds in EGFR using LigandFit.** (A) Iressa, (B) 2-O-Caffeoyl tartaric acid, (C) Emetine, (D) Rosmaricine, and (E) 2-O-Feruloyl tartaric acid. Binding site amino acids are shown in red (negatively charged amino acids) and blue (positively charged amino acids). Hydrogen bonds are color coded based on bond distance: 1.6–1.8Å (orange), 1.8–2.0Å (yellow), 2.0–2.2Å (light green), 2.2–2.4Å (neon green), and 2.4–2.5Å (dark green).
doi:10.1371/journal.pcbi.1002189.g002

Correlation between the predicted and observed $pIC_{50}$ activities on EGFR ligands of known activity using the constructed MLR and SVM models were illustrated in Figure 3a and 3b, respectively. Correlation coefficients based on the training set were 0.7858 for the MLR model and 0.8754 for the SVM model. Activity predictions of Iressa® and the TCM candidates using MLR and SVM were summarized in Table 1. Both models indicate that Iressa and the TCM candidates are compounds with acceptable predicted activities. Predicted activities ($pIC_{50}$) of Iressa by the trained MLR and SVM models were 6.715 and 5.110, respectively. The Iressa activity predicted by SVM was closer to experimentally determined Iressa activities ($pIC_{50}$) between 4.76–5.96 [57], thus SVM values may be more accurate predictions of the actual activity.

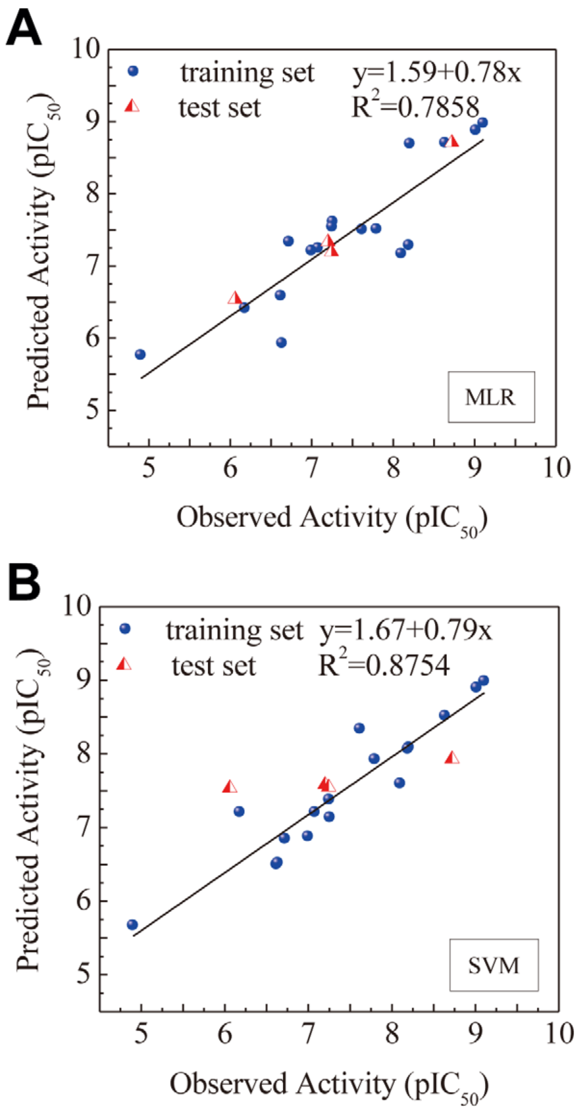

**Figure 3. Correlation of observed and predicted activity ($pIC_{50}$) using 2D-QSAR models.** (A) MLR and (B) SVM.
doi:10.1371/journal.pcbi.1002189.g003

**Table 2.** CoMFA and CoMSIA models as a factor of various fractions and the corresponding validation values.

| | Cross validation | | Non-cross validation | | | Fraction | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ONC | $q^2_{cv}$ | $r^2$ | SEE | F | S | H | D | A |
| **CoMFA** | 6 | 0.721 | 0.986 | 0.142 | 117.843 | 1.00 | - | - | - |
| **CoMFA** | | | | | | | | | |
| S | 6 | 0.764 | 0.975 | 0.189 | 65.257 | 1.00 | - | - | - |
| H | 6 | 0.331 | 0.980 | 0.168 | 83.249 | - | 1.00 | - | - |
| D | 6 | 0.236 | 0.945 | 0.281 | 28.780 | - | - | 1.00 | |
| A | 6 | −0.344 | 0.784 | 0.558 | 6.034 | - | - | - | 1.00 |
| SH | 6 | 0.541 | 0.986 | 0.141 | 118.218 | 0.37 | 0.63 | - | - |
| SD | 6 | 0.615 | 0.982 | 0.159 | 92.965 | 0.48 | - | 0.52 | - |
| SA | 6 | 0.718 | 0.977 | 0.180 | 72.008 | 0.53 | - | - | 0.47 |
| HD | 6 | 0.569 | 0.982 | 0.160 | 91.804 | - | 0.56 | 0.44 | - |
| HA | 6 | 0.265 | 0.979 | 0.176 | 76.040 | - | 0.60 | - | 0.40 |
| DA | 6 | 0.003 | 0.956 | 0.250 | 36.635 | - | - | 0.60 | 0.40 |
| SHD* | 6 | 0.662 | 0.988 | 0.133 | 134.272 | 0.26 | 0.40 | 0.35 | - |
| SHA | 6 | 0.441 | 0.983 | 0.156 | 96.519 | 0.26 | 0.42 | - | 0.33 |
| SDA | 6 | 0.664 | 0.979 | 0.173 | 78.179 | 0.33 | - | 0.38 | 0.30 |
| HDA | 6 | 0.543 | 0.979 | 0.175 | 76.649 | - | 0.42 | 0.32 | 0.26 |
| SHDA | 6 | 0.665 | 0.985 | 0.148 | 107.106 | 0.20 | 0.31 | 0.27 | 0.22 |

ONC: Optimal number of components.
SEE: Standard error of estimate.
F: F-test value.
PLS: partial least squares.
S: Steric.
H: Hydrophobic.
D: Hydrogen bond donor.
A: Hydrogen bond acceptor.
*: CoMISA model selected for 3D-QSAR simulation.
doi:10.1371/journal.pcbi.1002189.t002

## 3D QSAR Model

The results of CoMFA and CoMSIA model generation are detailed in Table 2. Steric field was the sole factor in the CoMFA model since the electrostatic field value was zero. Cross-validated ($q^2$) and non-cross-validated ($r^2$) correlation coefficient values of 0.721 and 0.986, respectively, indicated a high level of confidence for this model. The small standard error of estimates (SEE) and large F-test value further supported the reliability of this model. In contrast, CoMSIA models were influenced by multiple factors including steric field, hydrophobic region, and hydrogen bond donor/acceptors. Among all generated versions of the CoMSIA model, CoMSIA_SHD had the highest $r^2$ (0.988), lowest SEE (0.133), and highest F value (134.272), thus was selected as the optimum CoMSIA model for use in this study. The $pIC_{50}$ of 20 ligands predicted by the constructed CoMFA and CoMSIA models were compared with observed $pIC_{50}$ reported by Fidanze et al. [53] in Table 3. In general, both models gave similar predicted values and were close to the experimentally determined activities. Correlations between predicted and observed $pIC_{50}$ using CoMFA and CoMSIA models are summarized in Figure 4a and 4b, respectively. High correlation coefficients validated the reliability of the constructed CoMFA ($r^2 = 0.9860$) and CoMSIA($r^2 = 0.9877$) models.

Ligand activities of Iressa® and the TCM candidates can be predicted based on structural conformation to the 3D-QSAR feature map, including features in steric field, hydrophobic field, and H-bond donor/acceptor characteristics. As illustrated in Figure 5, Iressa and the TCM candidates were able to match the generated 3D-QSAR model features. The benzene in Iressa® favored steric and hydrophobic fields, and H-bond was favored between its amine group and Asp855. In 2-O-Caffeoyl tartaric acid, the benzene structure favored steric and hydrophobic fields, and the carboxyl group favored H-bond formation with Lys716 and Lys728. The carbon chain structure in Emetine contoured to the steric and hydrophobic fields, and the amine group favored H-bond formation with Asp855. Rosmaricine had benzene and isopropyl structures that favored steric and hydrophobic fields, and an amine group that favored H-bond with Asp 855. The benzene structure in 2-O-feruloyl tartaric acid favored steric fields and the carboxyl group favored H-bond formations with Lys716 and Lys728. Iressa® and the TCM candidates have structural components that contour to the features of the 3D-QSAR model, thus were likely to be biologically active.

**Table 3.** Observed and predicted activities of EGFR ligands using the constructed CoMFA and CoMSIA models.

| | | CoMFA | | CoMSIA | |
|---|---|---|---|---|---|
| Comp. | Observed $pIC_{50}^{*}$ | Predicted $pIC_{50}$ | Residual | Predicted $pIC_{50}$ | Residual |
| 1 | 6.620 | 6.571 | 0.049 | 6.600 | 0.020 |
| 2 | 7.081 | 7.192 | −0.111 | 7.230 | −0.149 |
| 3 | 7.260 | 7.234 | 0.026 | 7.147 | 0.113 |
| 4 | 6.638 | 6.394 | 0.244 | 6.522 | 0.116 |
| 5 | 8.102 | 8.337 | −0.235 | 8.275 | −0.173 |
| 6 | 8.721 | 8.508 | 0.213 | 8.493 | 0.228 |
| 7 | 6.060 | 5.940 | 0.120 | 6.012 | 0.048 |
| 8 | 6.180 | 6.237 | −0.057 | 6.247 | −0.067 |
| 9 | 7.000 | 6.893 | 0.107 | 6.952 | 0.048 |
| 10 | 6.721 | 6.828 | −0.107 | 6.717 | 0.004 |
| 11 | 7.201 | 7.293 | −0.092 | 7.322 | −0.121 |
| 12 | 8.208 | 8.149 | 0.059 | 7.806 | 0.402 |
| 13 | 9.108 | 9.077 | 0.031 | 9.167 | −0.059 |
| 14 | 9.018 | 9.059 | −0.041 | 9.023 | −0.005 |
| 15 | 8.638 | 8.563 | 0.075 | 8.566 | 0.072 |
| 16 | 7.252 | 6.377 | 0.875 | 6.012 | 1.240 |
| 17 | 7.244 | 7.210 | 0.034 | 7.159 | 0.085 |
| 18 | 7.796 | 7.790 | 0.006 | 7.710 | 0.086 |
| 19 | 7.620 | 7.744 | −0.124 | 7.729 | −0.109 |
| 20 | 8.194 | 8.089 | 0.105 | 8.216 | −0.022 |

*: Experimental values of ligand bioactivity adapted from Ref [53].
doi:10.1371/journal.pcbi.1002189.t003

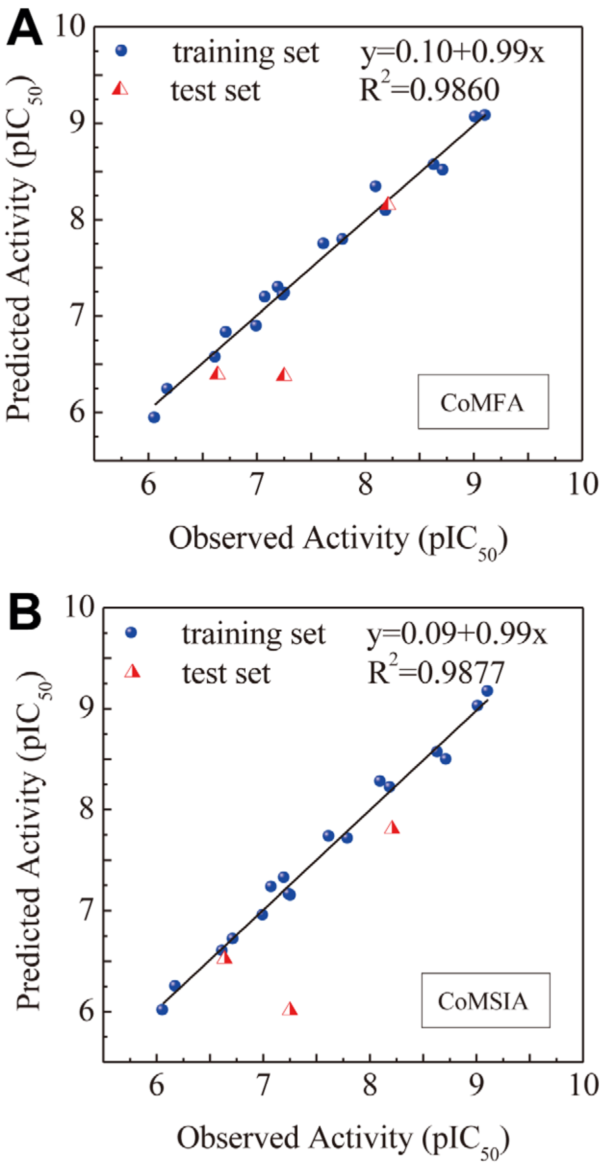

**Figure 4. Correlation of observed and predicted activity ($pIC_{50}$) using 3D-QSAR models.** (A) CoMFA and (B) CoMSIA.
doi:10.1371/journal.pcbi.1002189.g004

## Molecular Dynamics Simulation

Binding stability of the control and TCM candidates was validated using MD simulation. RMSDs of protein-ligand complex (Figure 6a) and individual ligand (Figure 6b) stabilized after 10 ns. The RMSDs of the protein-ligand complexes stabilized at approximately 1.6Å. With regard to individual ligands, the RMSDs of Iressa and 2-O-caffeoyl tartaric acid was 2.0 and 1.6Å, respectively. All other compounds registered RMSD values of approximately 1.0Å. The lower RMSD values of the TCM candidates suggest more stability within the receptor compared to Iressa. The energy trajectory of each compound is shown in Figure 6c. Complexes formed by Rosmaricine and 2-O-feruloyl tartaric acid had the lowest total energy (<−14,800 kcal/mol), followed by Iressa® and Emetine (approximately −14,700 kcal/mol), and 2-O-caffeoyl tartaric acid (−14,600 kcal/mol). Stabilization of total energy in ligand-protein complexes was achieved after 12 ns.

H-bond distance profiles in the EGFR receptor were summarized in Figure 7. A single H-bond between the amine group on Iressa® and the carboxyl group on Asp855 was formed after 9.74 ns and stabilized after 20 ns (Figure 7a). Two H-bonds were formed between the carboxyl group of 2-O-caffeoyl tartaric acid and Lys716 and Lys728 of the EGFR receptor (Figure 7b). The formation of two H-bonds contributed to a higher stability between 2-O-caffeoyl tartaric acid and the EGFR receptor. However, an increase in H-bond distance was observed towards the end of the 20 ns simulation period, suggesting a weakening of the H-bond at Lys728. Emetine formed a total of four H-bonds with the receptor, two with Asp722 and two with Ala855 (Figure 7c). Bond distances stabilized after 10 ns for Ala722 and 4 ns for Asp855. Rosmaricine formed three H-bonds each at Asp841 and Arg855 (Figure 7d). The multiple H-bonds enabled Rosmaricine to remain in a stable state within the protein. 2-O-Feruloyl tartaric acid also formed multiple H-bond at Lys716 and Lys728, enhancing its stability within the receptor site (Figure 7e). However, similar to 2-O-caffeoyl tartaric acid, an increase in H-bond distance was also observed at Lys728 for 2-O-feruloyl tartaric acid. These observations suggest that the bond at Lys728 weakens throughout the MD simulation process, and that the H-bond at Lys716 may be the primary bond for 2-O-caffeoyl tartaric acid and 2-O-feruloyl tartaric acid. In addition, periodic fluctuations in H-bond distances were observed in 2-O-caffeoyl tartaric acid, Rosmaricine, and 2-O-feruloyl tartaric acid. These phenomena can be attributed to the rotation of the amine group where the H-bond is formed. These

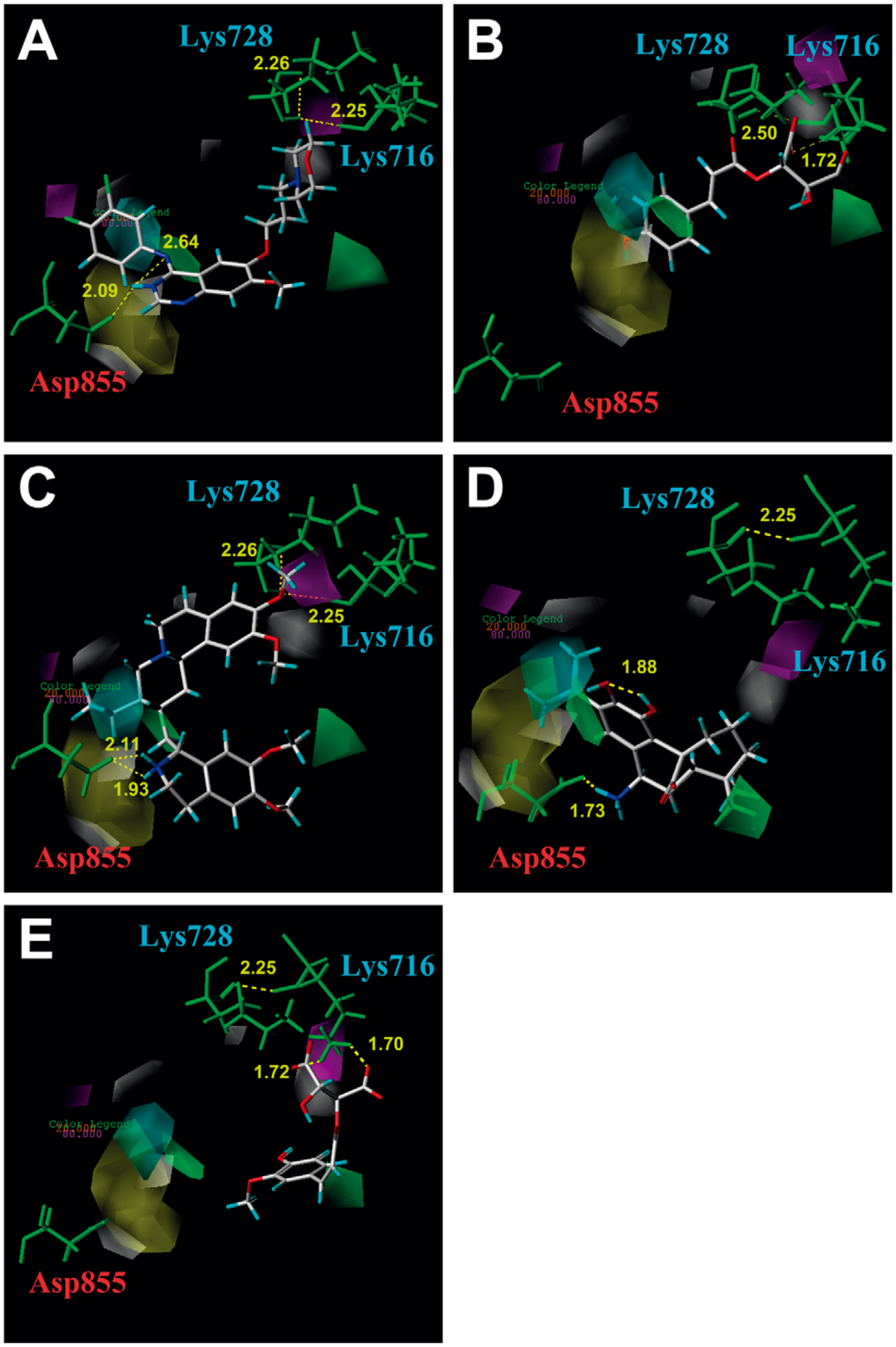

**Figure 5. Structural contouring of different compounds to 3D-QSAR mapping.** (A) Iressa, (B) 2-O-Caffeoyl tartaric acid, (C) Emetine, (D) Rosmaricine, and (E) 2-O-Feruloyl tartaric acid. 3D-QSAR features are represented by the following: steric field favor/disfavor (green/yellow), hydrophobic field favor/disfavor (cyan/white), and hydrogen bond donor avor/disfavor (magenta/orange). Bond distances are shown in yellow.
doi:10.1371/journal.pcbi.1002189.g005

MD results support our docking findings which identify Asp855, Lys716, and Lys 728 as key residues for docking.

As determined in the CoMSIA model, hydrophobic interactions were key factors contributing to ligand bioactivity. Toward the final 20 ns of analysis, hydrophobic amino acids surrounding the docking region were Leu718, Val726, Ala743, Cys775, Phe795, Cys797, and Leu844. The hydrophobic subgroups of Iressa®, Emetine, and Rosmaricine were surrounded by Val726, Cys797, and Leu844 (Figure 8a). Hydrophobic groups of 2-O-caffeoyl tartaric acid were also surrounded Val726, Cys797, and Leu844 (Figure 8b). The hydrophobic region of 2-O-feruloyl tartaric acid was attracted to the Phe795 on EGFR (Figure 8b). The significance of matching the hydrophobic region of the ligand to that of the receptor may be to increase stability of the

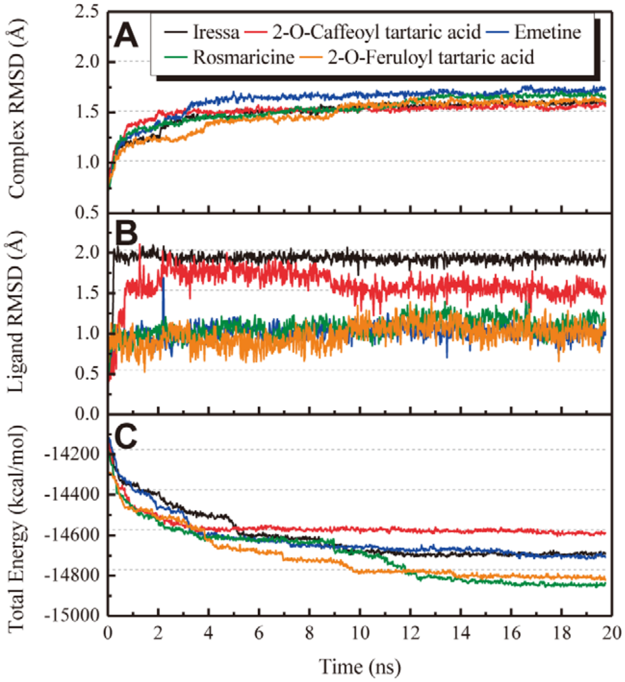

**Figure 6. RMSD and total energy during MD simulation.** (A) Protein-ligand complex RMSD (Å), (B) ligand RMSD (Å) and (C) total energy of protein-ligand complex.
doi:10.1371/journal.pcbi.1002189.g006

ligand-protein complex, and contribute to the bioactivity of the activated ligand. Our results indicate that Iressa® and the TCM candidates remained stable within the EFGR hydrophobic area following MD simulations.

## Conclusion

Structural and ligand based methods supported 2-O-caffeoyl tartaric acid, Emetine, Rosmaricine, and 2-O-feruloyl tartaric acid as potential EGFR inhibitors. Structurally, the TCM candidates were capable of forming H-bonds with key residues Asp855, Lys716, and Lys728 and matched hydrophobic regions of the receptor. Bioactivity of the candidates were evaluated using validated MLR, SVM, CoMFA, and CoMSIA models. All models indicated that the TCM candidates have good predicted bioactivity. Molecular simulation results further supported the high potential for the TCM candidates in drug development. Iressa®, the drug currently used clinically, bound to the ERGF receptor through a single H-bond at Asp855. In comparison, multiple H-bonds formed at Asp855 and additional H-bonds formed at Ala722 and Arg841 increase the stability of Emetine and Rosmaricine, respectively. The ability of carboxyl groups in 2-O-caffeoyl tartaric acid and 2-O-feruloyl tartaric acid to form multiple H-bond networks that directly blocked the ATP binding site was also a unique characteristic worthwhile of further investigation. Contour to hydrophobic regions of the TCM candidates within the receptor site provides additional support for the stability of the protein-ligand complex. In summary, using different simulation and validation methods, we have identified four TCM compounds that may have potential as novel EGFR inhibitors. As the four TCM compounds have two distinctive types of binding locations and bond formation within the EGFR binding site, we suggest exploring the possibility of connecting Emetine/Rosmaricine with 2-O-caffeoyl tartaric acid/2-O-feruloyl tartaric acid through a spacer. The connection could allow more of points of attachment, which in turn would contribute to more stable binding within the tyrosine kinase site.

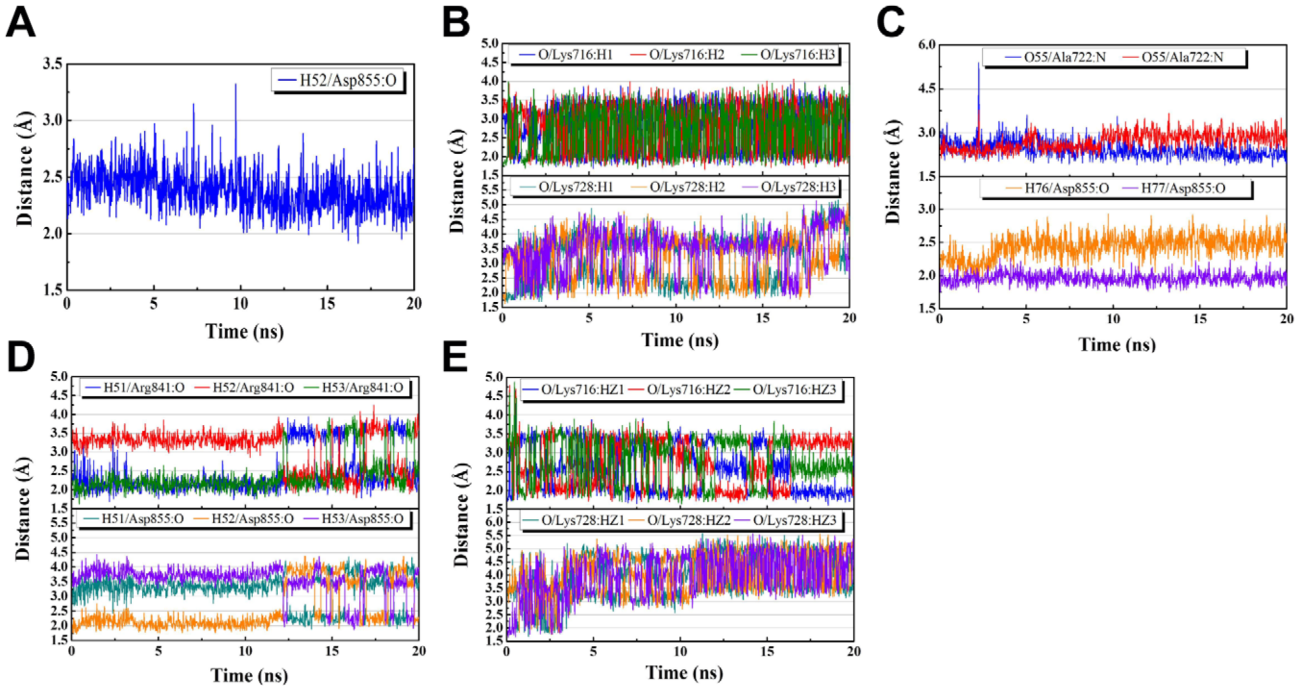

**Figure 7. Hydrogen bond distance profile during MD simulation.** (A) Iressa, (B) 2-O-Caffeoyl tartaric acid, (C) Emetine, (D) Rosmaricine, (E) 2-O-Feruloyl tartaric acid.
doi:10.1371/journal.pcbi.1002189.g007

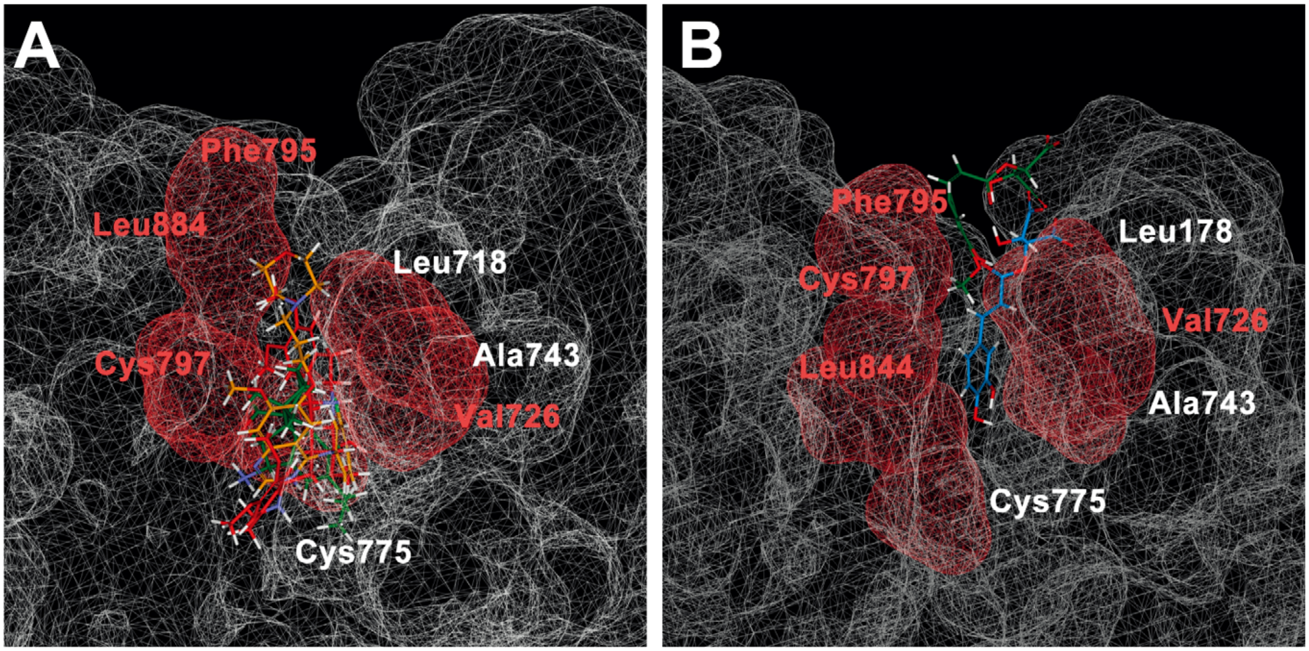

**Figure 8. Hydrophobic interactions of different compounds in EGFR.** (A) Iressa (orange), Emetine (green), and Rosmaricine (violet) (B) 2-O-Caffeoyl tartaric acid (blue), and 2-O-feruloyl tartaric acid (green). The hydrophobic regions in the binding site are depicted in red, and specific hydrophobic amino acids close to TCM compounds are indicated in red.

doi:10.1371/journal.pcbi.1002189.g008

## Supporting Information

**Table S1 Molecular structures and biological activities of ligands used for model training.** Structural details of the 20 ligands adopted for ligand-based studies are listed within this table. Experimental bioactivity values for each ligand were adapted from Ref [53].
(DOC)

## Acknowledgments

We are grateful to the Asia University cloud-computing facilities.

## Author Contributions

Conceived and designed the experiments: CYCC. Performed the experiments: SCY HYC CYCC. Analyzed the data: SCY SSC. Contributed reagents/materials/analysis tools: HYC CYCC. Wrote the paper: SCY SSC CYCC.

## References

1. de Oliveira AN, Bocca CC, Carvalho JE, Ruiz AL, Silva TP, et al. (2010) New substituted 4-arylaminoquinazolines as potent inhibitors of breast tumor cell lines: *in vitro* and docking experiments. Eur J Med Chem 45: 4339–4342.
2. El-Azab AS, Al-Omar MA, Abdel-Aziz AA, Abdel-Aziz NI, el-Sayed MA, et al. (2010) Design, synthesis and biological evaluation of novel quinazoline derivatives as potential antitumor agents: molecular docking study. Eur J Med Chem 45: 4188–4198.
3. Wang P, Hu L, Liu G, Jiang N, Chen X, et al. (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS One 6: e18476.
4. Herbst RS (2004) Review of epidermal growth factor receptor biology. Int J Radiat Oncol Biol Phys 59: 21–26.
5. Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1: 1–17.
6. Franklin WA, Veve R, Hirsch FR, Helfrich BA, Bunn PA, Jr. (2002) Epidermal growth factor receptor family in lung cancer and premalignancy. Semin Oncol 29: 3–14.
7. Linggi B, Carpenter G (2006) ErbB receptors: new insights on mechanisms and biology. Trends Cell Biol 16: 649–656.
8. Chan TO, Rittenhouse SE, Tsichlis PN (1999) AKT/PKB and other D3 phosphoinositide-regulated kinases: kinase activation by phosphoinositide-dependent phosphorylation. Ann Rev Biochem 68: 965–1014.
9. Either SP (2002) Signal transduction pathways: the molecular basis for targeted therapies. Semin Radiat Oncol 12: 3–10.
10. Shibata T, Kawano T, Nagayasu H, Okumura K, Arisue M, et al. (1996) Enhancing effects of epidermal growth factor on human squamous cell carcinoma motility and matrix degradation but not growth. Tumour Biol 17: 168–175.
11. Petit AM, Rak J, Hung MC, Rockwell P, Goldstein N, et al. (1997) Neutralizing antibodies against epidermal growth factor and ErbB-2/neu receptor tyrosine kinases down-regulate vascular endothelial growth factor production by tumor cells in vitro and in vivo: angiogenic implications for signal transduction therapy of solid tumors. Am J Pathol 151: 1523–1530.
12. Mitsudomi T, Yatabe Y (2010) Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer. FEBS J: 277: 301–308.
13. Sordella R, Bell DW, Haber DA, Settleman J (2004) Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. Science 305: 1163–1167.
14. Wakeling AE, Barker AJ, Davies DH, Brown DS, Green LR, et al. (1996) Specific inhibition of epidermal growth factor receptor tyrosine kinase by 4-anilinoquinazolines. Breast Cancer Res Treat 38: 67–73.
15. Merlo V, Longo M, Novello S, Scagliotti GV (2011) EGFR pathway in advanced non-small cell lung cancer. Front Biosci 3: 501–517.
16. Van den Eynde M, Baurain JF, Mazzeo F, Machiels JP (2011) Epidermal growth factor receptor targeted therapies for solid tumours. Acta Clin Belg 66: 10–17.
17. Sherman SI (2011) Targeted therapies for thyroid tumors. Mod Pathol 24: S44–52.
18. Medina PJ, Goodin S (2008) Lapatinib: a dual inhibitor of human epidermal growth factor receptor tyrosine kinases. Clin Ther 30: 1426–1447.
19. Johnston SR, Leary A (2006) Lapatinib: a novel EGFR/HER2 tyrosine kinase inhibitor for cancer. Drugs Today (Barc) 42: 441–453.
20. Forsythe B, Faulkner K (2004) Overview of the tolerability of gefitinib (IRESSA) monotherapy: clinical experience in non-small-cell lung cancer. Drug Saf 27: 1081–1092.
21. AustraZeneca Pharmaceuticals (2011) Iressa Product Information. Available: http://www.iressa.com/product-information. Accessed 15 May 2011.
22. Chen L, Huang T, Shi XH, Cai YD, Chou KC (2010) Analysis of protein pathway networks using hybrid properties. Molecules 15: 8177–8192.
23. Huang T, Shi XH, Wang P, He Z, Feng KY, et al. (2010) Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks. PLoS One 5: e10972.

24. Chou KC (1996) Review: Prediction of HIV protease cleavage sites in proteins. Anal Biochem 233: 1–14.
25. Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246–255.
26. Shen HB, Chou KC (2008) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal Biochem 375: 388–390.
27. Chou KC, Wu ZC, Xiao X (2011) iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6: e18258.
28. Chou KC (2004) Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134.
29. Chou KC, Shen HB (2009) Review: Recent advances in developing web-servers for predicting protein attributes. Nat Sci 2: 63–92.
30. Liao QH, Gao QZ, Wei J, Chou KC (2011) Docking and molecular dynamics study on the inhibitory activity of novel inhibitors on epidermal growth factor receptor (EGFR). Med Chem 7: 24–31.
31. Mustafa M, Mirza A, Kannan N (2011) Conformational regulation of the EGFR kinase core by the juxtamembrane and C-terminal tail: a molecular dynamics study. Proteins 79: 99–114.
32. Krug M, Wichapong K, Erlenkamp G, Sippl W, Schachtele C, et al. (2011) Discovery of 4-benzylamino-substituted alpha-carbolines as a novel class of receptor tyrosine kinase inhibitors. ChemMedChem 6: 63–72.
33. Zheng Y, Wu X, Xue B, Li M, Ji M (2010) Design, synthesis, docking and antitumor activity of quinazolino [3, 4-a] thieno [3, 2-d] pyrimidin-8-one derivatives. Chem Biol Drug Des 76: 285–290.
34. Qian Y, Zhang HJ, Zhang H, Xu C, Zhao J, et al. (2010) Synthesis, molecular modeling, and biological evaluation of cinnamic acid metronidazole ester derivatives as novel anticancer agents. Bioorg Med Chem 18: 4991–4996.
35. Lv PC, Li HQ, Sun J, Zhou Y, Zhu HL (2010) Synthesis and biological evaluation of pyrazole derivatives containing thiourea skeleton as anticancer agents. Bioorg Med Chem 18: 4606–4614.
36. Ali HI, Nagamatsu T, Akaho E (2011) Structure-based drug design and AutoDock study of potential protein tyrosine kinase inhibitors. Bioinformation 5: 368–374.
37. Nandi S, Bagchi MC (2010) 3D-QSAR and molecular docking studies of 4-anilinoquinazoline derivatives: a rational approach to anticancer drug design. Mol Divers 14: 27–38.
38. Cheng WY, Wu SL, Hsiang CY, Li CC, Lai TY, et al. (2008) Relationship between San-Huang-Xie-Xin-Tang and its herbal components on the gene expression profiles in HepG2 cells. Am J Chin Med 36: 783–797.
39. Liu SH, Liu YF, Liou SH, Lin YL, Hsiao YC, et al. (2010) Mortality and cancer incidence among physicians of traditional Chinese medicine: a 20-year national follow-up study. Occup Environ Med 67: 166–169.
40. Hsu CH, Yu MC, Lee CH, Lee TC, Yang SY (2003) High eosinophil cationic protein level in asthmatic patients with "heat" Zheng. Am J Chin Med 31: 277–283.
41. Huang HJ, Lee KJ, Yu HW, Chen CY, Hsu CH, et al. (2010) Structure-based and ligand-based drug design for HER2 receptor. J Biomol Struct Dyn 28: 23–37.
42. Chen CY, Chen CYC (2010) Insights into designing the dual-targeted HER2/HSP90 inhibitors. J Mol Graph Model 29: 21–31.
43. Choowongkomon K, Sawatdichaikul O, Songtawee N, Limtrakul J (2010) Receptor-based virtual screening of EGFR kinase inhibitors from the NCI diversity database. Molecules 15: 4041–4054.
44. Abou-Seri SM (2010) Synthesis and biological evaluation of novel 2,4′-bis substituted diphenylamines as anticancer agents and potential epidermal growth factor receptor tyrosine kinase inhibitors. Eur J Med Chem 45: 4113–4121.
45. Pasha FA, Muddassar M, Srivastava AK, Cho SJ (2010) In silico QSAR studies of anilinoquinolines as EGFR inhibitors. J Mol Model 16: 263–277.
46. Coumar MS, Chu CY, Lin CW, Shiao HY, Ho YL, et al. (2010) Fast-forwarding hit to lead: aurora and epidermal growth factor receptor kinase inhibitor lead identification. J Med Chem 53: 4980–4988.
47. Konkimalla VB, Efferth T (2010) Inhibition of epidermal growth factor receptor-overexpressing cancer cells by camptothecin, 20-(N,N-diethyl) glycinate. Biochem Pharmacol 80: 39–49.
48. Chen CYC (2011) TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS One 6: e15939.
49. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273: 236–247.
50. The UniProt Consortium (2010) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39: D214–D219.
51. Yun CH, Boggon TJ, Li Y, Woo MS, Greulich H, et al. (2007) Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. Cancer Cell 11: 217–227.
52. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, et al. (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30: 1545–1614.
53. Fidanze SD, Erickson SA, Wang GT, Mantei R, Clark RF, et al. (2010) Imidazo[2,1-b]thiazoles: multitargeted inhibitors of both the insulin-like growth factor receptor and members of the epidermal growth factor family of receptor tyrosine kinases. Bioorg Med Chem Lett 20: 2452–2455.
54. Picard RR, Cook RD (1984) Cross-validation of regression models. J Am Stat Assoc 79: 575–583.
55. Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. J Mach Learn Res 6: 1889–1918.
56. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. Anal Biochem 370: 1–16.
57. Pernas FG, Allen CT, Winters ME, Yan B, Friedman J, et al. (2009) Proteomic Signatures of Epidermal Growth Factor Receptor and Survival Signal Pathways Correspond to Gefitinib Sensitivity in Head and Neck Cancer. Clin Cancer Res 15: 2361–2372.