Supplementary Material (ESI) for Medicinal Chemistry Communications


Supporting Information

# Computational ligand-based rational design:

# Role of conformational sampling and force fields in model development

Jihyun Shim and Alexander D. MacKerell, Jr.

Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland 21201

Table S1. Overview of statistical methods used in 2D-QSAR

| Method | Description |
|---|---|
| Stepwise selection | Descriptors are added one by one (forward selection) or deleted (backward selection) to find significant descriptors yielding the best statistics. |
| Replacement method | Initially, a subset of descriptors are selected randomly and iteratively one of them is replaced with one from the other set. Heuristics avoiding full searching increases efficiency but there is possibility of being trapped in local minima. |
| Genetic algorithm (GA) | In GA a model is represented as a chromosome and descriptors are genes on the chromosome. During reproduction, chromosomes undergo mutations and recombination generating diverse descendant chromosomes. Descriptors in high scoring chromosomes or frequently appearing descriptors in the entire population are deemed important. |
| Multiple linear regression (MLR) | Coefficients ($a_n$) are determined in the equation, "Activity = $a_1\text{Descriptor}_1+a_2\text{Descriptor}_2+\ldots+a_n$" by least square fitting often using the Levenberg-Marquardt algorithm[1]. |
| Principal component regression (PCR) | Eigenvectors of the covariance matrix of descriptors are used as independent variables in the regression. These orthogonal vectors (principal components) describe the direction of maximum variance in descriptor space, resulting in grouping a large number of descriptors in the final models. |
| Partial least square regression (PLR) | A process to reduce the number of variables by finding principal components (or latent variable) as in PCR, PLS includes correlation with dependent variables. Therefore, the maximum variances reflect both descriptor space and activity space. |
| Linear discriminant analysis (LDA) | LDA performs linear transformation of descriptors to better discriminate the categorical data by minimizing within-class variance and maximizing between-class variance. Solution is found based on Bayes theorem. |
| Support vector machine (SVM) | SVM trains a model to find a hyperplane of descriptors by separating data into subsets with maximum margins. Vectors on the margins are called support vectors and they are components of a kernel function, which is used in mapping data into a new dimension. It was expanded for non-linear classification and regression by mapping input vectors into higher dimensions. |
| Decision tree (DC) | Training set data are recursively partitioned and pruned based on the best splitting descriptors from the top node to the end nodes in the binary tree. |

| | |
|---|---|
| Random forest (RF) | As an ensemble of DC, each tree votes for the activities. Individual trees are grown by using a randomly selected subset of full descriptor set. |
| K-nearest neighbor[2] | Data is divided into training and test set. Each molecule in the test set is classified according to the majority among k-nearest neighbors and decision surface is learned. |
| Artificial neural network (feed forward network) [3, 4] | Descriptors in the input nodes are connected with nodes in hidden layers and weights on the nodes are trained to produce activity in an output node. |
| Self organizing map (SOM, Kohonen map) | In contrast to a feed forward network, a self-organizing map is subjected to unsupervised learning and input nodes are projected to nodes in a rectangular forms (feature map) with weights. Training is done by minimizing distances between nodes and the result of training is clusters or organized patterns in the feature map. |
| Adaptive fuzzy partition algorithm | It aims to find the best descriptors by splitting data sets by fuzzy rules. Fuzzy logic is originated from human reasoning making a correct judgment based on uncertain information. Models are trained by a set of adaptive IF-THEN rules and fuzzy scoring function. |
| Gene expression programming (GEP) | While being similar to GA, uniqueness of GEP is using expression of chromosome and fitness function to evaluate the phenotype. From a chromosome, different genes (descriptors) may be expressed according to reading frames resulting diversification of child chromosomes. |
| Gaussian processes (GP) | While other nonlinear regressions have fixed function and varying parameters (weights) during model development, GP uses varying Gaussian functional forms that are trained by Bayesian inference. |
| Project pursuit regression (PPR) | To reduce the problem of increasing volume of higher dimensional space, without introducing higher dimensions, PPR projects input data into 1D space as SOM does. A series of transformations are trained to explain activities. |
| Local lazy regression | Local QSARs are generated in clusters of molecules and prediction of test compounds is done by searching nearest neighbors. |

Computational Methods

MD simulations (Figure 1) were performed using the program CHARMM[5, 6], with the CHARMM22/CMAP force field[6, 7] and the TIP3P water model[8]. Two MD simulations were done in the gas phase and in explicit solvent to observe the influence of the surrounding media on conformational sampling. For the gas phase simulation, Langevin dynamics[9] were used with a friction coefficient of 5 $ps^{-1}$ in the absence of water. Simulations were performed at 300K with the equations of motion integrated using the Leap-Frog integrator[10] every 1fs for a total of 10ns with coordinates saved every 0.5ps for analysis. Covalent bonds involving hydrogen atoms were constrained to their equilibrium bond length by the SHAKE algorithm[11]. In the explicit solvent MD simulation, Leu-Enkephalin was immersed in a 32 Å cubic water box and waters with the oxygen within 2.8 Å of the Leu-Enkephalin deleted resulting in 976 water molecules. Periodic boundary conditions[9] were used in the solvent simulations and the nonbond interactions were truncated at 12 Å with smoothing of the Lennard-Jones interactions from 10 Å by a switching function[12] and the nonbond pair list was generated out to 16 Å. A long-range correction[13] was used to account for LJ interactions beyond the cutoff distance while long-range electrostatic interactions were calculated using the particle mesh Ewald method[14]. The system was simulated in the NPT ensemble (300K, 1atm) using the Hoover thermostat and Langevin piston[15-18] to control the pressure with a mass of 400 amu and collision frequency of 20 $ps^{-1}$. The Hamiltonian replica exchange (HREMD)[19, 20] simulation was performed using the REPDSTR module in CHARMM which enables replicas to read in different FF parameters. Perturbation of the Hamiltonian was performed using CMAP utility[7] by modifying the ($\varphi$,$\psi$) potential energy surface. The default CHARMM22/CMAP was considered as the $\lambda$=0 state ($CMAP_{\lambda=0}$) and as the fully perturbed state ($CMAP_{\lambda=1}$) a "flat" ($\varphi$,$\psi$) energy surface was used (ie. the change in energy as a function of $\varphi$,$\psi$ was zero). Perturbations of $\lambda$=0.14, 0.19, 0.27, 0.37, 0.52,

0.72 were used between $CMAP_{\lambda=0}$ and $CMAP_{\lambda=1}$ and exchanges were attempted every 0.5ps between adjacent replicas. HREMD simulations were carried out on the same solvated Leu-Enkephalin used for the standard MD simulation and the same dynamics settings were used. Conformations from the $\lambda=0$ state replica alone were subjected to analysis.

For the conformational analysis of Leu-Enkephalin, the distance was measured between the aromatic ring in Tyr (pharmacophoric point A) and that in Phe (pharmacophoric point B). Angle ABN, where pharmacophoric point N is N terminal nitrogen, was calculated and it was combined with distance AB to produce the 2-D probability distribution. The distances and angles were obtained from all 20,000 conformations and the bin sizes for calculation of the probability densities were 0.1 Å and 1°.

References


1. K. Levenberg, *he Quarterly of Applied Mathematics*, 1944, **2**, 164-168.
2. T. Cover and P. Hart, *Information Theory, IEEE Transactions on*, 1967, **13**, 21-27.
3. S. Agatonovic-Kustrin and R. Beresford, *Journal of Pharmaceutical and Biomedical Analysis*, 2000, **22**, 717-727.
4. D. A. Winkler, *Molecular Biotechnology*, 2004, **27**, 139-168.
5. B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *Journal of Computational Chemistry*, 2009, **30**, 1545-1614.
6. A. D. J. MacKerell, B. Brooks, C. L. I. Brooks, L. Nilsson, B. Roux, Y. Won and M. Karplus, *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*, John Wiley & Sons: Chichester, 1998.
7. A. D. MacKerell, M. Feig and C. L. Brooks, *Journal of Computational Chemistry*, 2004, **25**, 1400-1415.
8. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926-926.
9. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, USA, 1989.
10. R. W. Hockney, *Methods Comput Phys*, 1970, **9**, 136-211.
11. J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *Journal of Computational Physics*, 1977, **23**, 327-341.
12. P. J. Steinbach and B. R. Brooks, *Journal of Computational Chemistry*, 1994, **15**, 667-683.
13. M. C. Pitman, F. Suits, A. D. MacKerell and S. E. Feller, *Biochemistry*, 2004, **43**, 15318-15328.
14. T. Darden, D. York and L. Pedersen, *The Journal of Chemical Physics*, 1993, **98**, 10089-10089.
15. H. C. Andersen, *The Journal of Chemical Physics*, 1980, **72**, 2384-2384.
16. S. E. Feller, Y. Zhang, R. W. Pastor and B. R. Brooks, *The Journal of Chemical Physics*, 1995, **103**, 4613-4613.
17. W. G. Hoover, *Physical Review A*, 1985, **31**, 1695-1695.
18. S. Nosé and M. L. Klein, *Molecular Physics: An International Journal at the Interface Between Chemistry and Physics*, 1983, **50**, 1055-1055.
19. Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, 1999, **314**, 141-151.
20. Y. Sugita and Y. Okamoto, *cond-mat/0009119*, 2000.