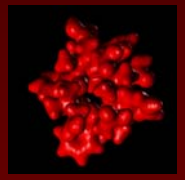
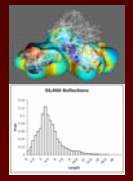
Advances in Protein Dissimilarity Assessment and Protein- Surface Interaction Modeling using Hybrid Shape/Electronic Property Descriptors and SVM Methods

C. Matthew Sundling, Qiong Luo and Curt M. Breneman* (brenec@rpi.edu), Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, NY 12180
Kristin P. Bennett, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180

Introduction
Property/Shape Encoding (PEST)

Methods for representing and encoding the electronic and shape properties of molecules have undergone a great deal of development over the past several years. It has now become common to consider the distribution of a variety of electronic functions on a molecular surface as a valuable means of characterizing its features. This is exemplified by the increasing use of polar or hydrophobic surface area descriptors as well as RECON electron density-based molecular surface descriptors in QSAR and QSPR modeling. When surface area histograms such as these are coupled with a shape-based encoding system such as Zauhar's "Shape Signatures" (Nagarajan et al., 2005), the resulting "PEST" (Property-Encoded Surface Translator) descriptors allow molecular shape and property information to be stored in an orientation-independent and alignment-free manner (Breneman et al., 2003). Recently, PEST technology has been adapted for use with protein surfaces, resulting in "PPEST" – a technique for characterizing the electronic and shape properties of the solvent-accessible surfaces of proteins.





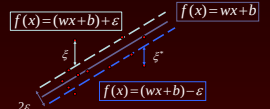
K. Nagarajan, R. Zauhar, and W.J. Welsh; *J. Chem. Inf. Model.*, **45,** 49 (2005).
C.M. Breneman, C. M. Sundling, N. Sukumar, L. Shen, W.P. Katt and M.J. Embrechts; *J. Computer-Aided Mol. Design* **17,** p.231 (2003).
W. Heiden, G. Moeckel, and J. Brickmann; *J. Comput.-Aided Mol. Design* **7,** 503 (1993).
E. Audry, J.-P. Dubost, J.-C. Colleter, and P. Dallet; *J. Med. Chem. Chim. Ther.* **21,** 71 (1986).
G. E. Kellogg, S. F. Semus, and D. J. Abraham; *J. Computer Aided Mol. Design*, **5,** 545 (1991).

Protein PEST (PPEST) Descriptors

The new PPEST method (Protein Property-Encoded Surface Translator) uses a technique akin to ray-tracing to explore the volume enclosed by the solvent-accessible surface of a protein. Ray-length distributions are derived from the ray-tracing results, (see histogram to the left), and can include joint dependence on a set of surface properties, such as molecular lipophilicity potentials (MLP) or molecular electrostatic potentials (EP) (see 2D distributions below). The 2D distributions, or "profiles", generated by PPEST can be used to assess the level of similarity between one protein and another, and as descriptors in classification or regression modeling.

In order to quantify lipophilic or hydrophilic potentials on a molecular surface, these local properties may be projected onto the solvent-accessible surface using local atomic lipophilicities together with one of several empirical mapping functions. One example by Heiden et al. (Heiden et al., 1993) involves the use of a 'molecular lipophilicity potential' (MLP) derived from their technique of Molecular Hydrophobic Mapping (MHM) that is based on a Fermi-type distance function (see Equation below). This hydrophobicity model is an adaptation of previous MLP models (hence MLP2) by Audry et al. (Audry et al., 1986) and takes into account the possibility that long-range distance dependency of the individual atomic potential contributions may lead to overcompensation of local effects. In this paradigm, atoms which are far away from the surface point do not contribute significantly to the local hydrophobicity. The HINT program (Kellogg et al., 1991) provides another approach to for analyzing lipophilic/hydrophilic properties on protein surface. Although the MLP functions are not based on rigorous physical concepts, they are useful for quantifying lipophilicity values on a molecular surface and generate reasonable molecular surfaces property maps.

Support Vector Regression



Minimize: Empirical error + Complexity

$$min.\quad C\sum_{i=1}^{n} L_\varepsilon(x_i) + \|w\|$$

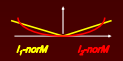
$$\begin{aligned} min.\quad & C\nu\varepsilon + b + \sum_{i=1}^{n}(u_i + v_i) + \frac{C}{l}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ & y_j - \sum_{i=1}^{n}(u_i - v_i)x_{ji} - b \le \varepsilon + \xi_j \\ s.t.\quad & \sum_{i=1}^{n}(u_i - v_i)x_{ji} + b - y_j \le \varepsilon + \xi_j^* \\ & u_i, v_i, \xi_j, \xi_j^*, \varepsilon \ge 0, \quad j = 1,\ldots,l \quad i = 1,\ldots,n \end{aligned}$$

- Linear hypotheses
- Empirical error
  - ε-insensitive loss function:

$$L_\varepsilon(x) = \max(0, |y - f(x)| - \varepsilon)$$

- Complexity control
  - l_1-norm weight vector:

$$\|w\|_1 = \sum_{i=1}^{n} |w_i|$$



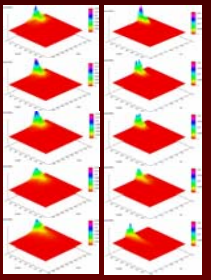
Support Vector Machine regression uses a kernel function to produce either a linear or non-linear model from a training dataset. In the case of protein interaction modeling, important descriptors are selected using one-norm SVM technology, and the final models are created using two-norm non-linear SVM regression. Results of this modeling effort are typically similar or better than those obtained using KPLS for protein interaction datasets.

The graphics shown below illustrate the improvement in SVM regression models created using RECON and MOE descriptors when PPEST descriptors are also used. In the first case, only RECON and MOE descriptors were used in the model, while in the second example, marked improvement was noted when PPEST shape/property descriptors were included. An important indication of the quality of the PPEST-derived model is the smaller number of features selected in the SVM regression model.

Protein Dissimilarity Measurement

PPEST shape/property data may be used to quantitatively compare shape-encoded surface distributions of either MLP2 or EP on different protein surfaces in an alignment-free manner. This technique provides a new method for classifying protein behavior. Presented below are two examples of the use of PPEST-based functions for assessing shape/property dissimilarity between protein surfaces:

$$D_{ij}^{A} = 1 - \frac{2\cdot\sum_{k=1}^{K}\min\left(x_{ik}, x_{jk}\right)}{\sum_{k=1}^{K}x_{ik} + \sum_{k=1}^{K}x_{jk}} \qquad D_{ij}^{B} = 1 - \frac{\sum_{k=1}^{K}\min\left(x_{ik}, x_{jk}\right)}{\sum_{k=1}^{K}\max\left(x_{ik}, x_{jk}\right)}$$



x_{ik} is the value of the k^{th} descriptor for the i^{th} protein

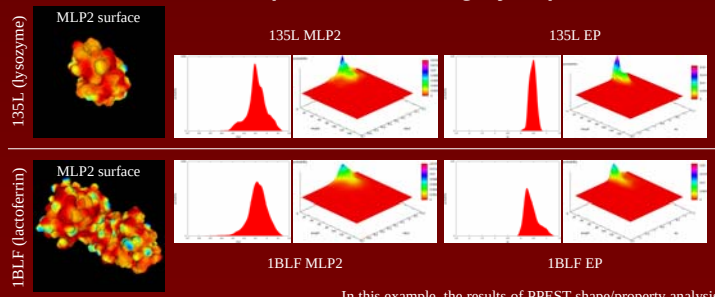
D ranges from 0 to 1

0 = identical

1 = nothing in common

The shape/property dissimilarity of PPEST distributions are presented in the table below. Note that 135L and 1BLF EP distributions are found to be similar, even though the size of the two proteins are significantly different.

Protein PEST Descriptors for Modeling Hydrophobic Interaction Chromatography

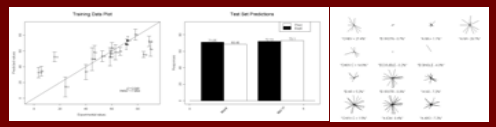


$$MLP2(d) = \frac{\sum_i f_i g(d_i)}{\sum_i g(d_i)} \qquad g(d_i) = \frac{e^{-a(d_i - d_{cut-off})} + 1}{e^{a(d_i - d_{cut-off})} + 1}$$

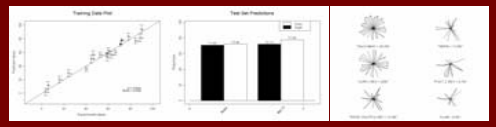
f_i is the partial lipophilicity of the i^{th} atom of a molecule and d_i is the distance of the surface point in 3D space from atom i, and where a proximity distance of $d_{cut-off}$ = 4Å and a = 1.5 are used.

In this example, the results of PPEST shape/property analysis of lysozyme and lactoferin are presented. The protein surfaces were derived using the MOE v. 2004.03 'GaussAccessible' surface function. The 1D distributions are histograms of MLP2/EP surface property distributions, while the 2D profiles are PPEST distributions that couple ray-tracing geometry with MLP2/EP surface property information.

P.PHENYL (RECON+MOE)



P.PHENYL (RECON+MOE+PPEST)



Summary

Electron Density-Derived shape/property descriptors may be used for modeling protein/surface interactions

Predictive models for protein chromatography can be built using Protein-PEST (PPEST) descriptors and SVM regression methods

Protein dissimilarity may be quantified using PPEST techniques