A New Method to Estimate Ligand-Receptor Energetics*

Joel R. Bock and David A. Gough‡

In the discovery of new drugs, lead identification and optimization have assumed critical importance given the number of drug targets generated from genetic, genomics, and proteomic technologies. High-throughput experimental screening assays have been complemented recently by "virtual screening" approaches to identify and filter potential ligands when the characteristics of a target receptor structure of interest are known. Virtual screening mandates a reliable procedure for automatic ranking of structurally distinct ligands in compound library databases. Computing a rank score requires the accurate prediction of binding affinities between these ligands and the target. Many current scoring strategies require information about the target three-dimensional structure. In this study, a new method to estimate the free binding energy between a ligand and receptor is proposed. We extend a central idea previously reported (Bock, J. R., and Gough, D. A. (2001) Predicting protein-protein interactions from primary structure. *Bioinformatics* 17, 455–460; Bock, J. R., and Gough, D. A. (2002) Whole-proteome interaction mining. *Bioinformatics*, in press) that uses simple descriptors to represent biomolecules as input examples to train a support vector machine (Smola, A. J., and Schölkopf, B. (1998) *A Tutorial on Support Vector Regression, Neuro-COLT Technical Report NC-TR-98-030*, Royal Holloway College, University of London, UK) and the application of the trained system to previously unseen pairs, estimating their propensity for interaction. Here we seek to learn the function that maps features of a receptor-ligand pair onto their equilibrium free binding energy. These features do not comprise any direct information about the three-dimensional structures of ligand or target. In cross-validation experiments, it is demonstrated that objective measurements of prediction error rate and rank-ordering statistics are competitive with those of several other investigations, most of which depend on three-dimensional structural data. The size of the sample (n = 2,671) indicates that this approach is robust and may have widespread applicability beyond restricted families of receptor types. It is concluded that newly sequenced proteins, or those for which three-dimensional crystal structures are not easily obtained, can be rapidly analyzed for their binding potential against a library of ligands using this methodology. *Molecular & Cellular Proteomics 1:904–910, 2002.*

The process of developing a new drug involves seven major steps (1). (i) First, a disease is identified, and then (ii) drug targets (usually proteins), the activation or inhibition of which is thought to alter the disease state, within the cell are hypothesized. Once targets are hypothesized, the next task is to (iii) identify potential lead compounds that will bind to the target. These leads are subsequently (iv) optimized with respect to their structural characteristics in the context of the target binding site and then subjected to (v) preclinical and (vi) clinical trials to determine their bioavailability and therapeutic potential. The final step is to (vii) optimize efficacy, toxicity, and pharmacokinetic properties. This may involve the use of pharmacogenomic techniques to tailor compounds to a subset of the patient population that is predisposed to a disease.

Pharmaceutical companies are exposed to great financial risk in the course of identifying viable drugs to treat a certain condition or disease. There are also tremendous direct and indirect (opportunity) costs associated with delaying the removal of non-viable drugs from this drug discovery "pipeline" until the latest stages of the process.

A huge number of drug targets have been generated from genetic, genomic, and proteomic technologies. Accordingly, the lead identification and optimization steps have assumed critical importance. High-throughput experimental screening assays (30) have been complemented recently by computational ("virtual screening") approaches to identify and filter potential ligands when the characteristics of the target receptor structure of interest are known (3, 35). In virtual screening, databases of compound libraries are searched, and scoring or discrimination functions are used to select the "best" candidate compounds for biological activity analysis (32).

The scoring of ligands in virtual screening is often associated with computational docking simulations that mate receptor and cognate small molecule ligand in three-dimensional space. To provide broad generalization in "chemical diversity" space, computing this score requires the accurate prediction of binding affinities of many structurally distinct ligands (31). Three main methodologies have been identified for free binding energy calculations. In order of computational complexity, these are: 1) knowledge-based scoring functions, 2) partitioning the binding energy into biophysical energy terms, and 3)

molecular dynamics (7). The most accurate computations are represented by molecular dynamic techniques, but their inherent computational intensity precludes their application to industrial-size chemical databases.

Regression-based scoring functions, as exemplified by the work of Böhm (8), are fast but require a three-dimensional structure of the receptor. This prohibits their use in cases where the structure is difficult to obtain, such as with transmembrane proteins. The accuracy of such methods has also been called into question. A recent investigation concluded that "no significant correlation" existed between Böhm-type scores and experimentally determined binding affinities for a group of 15 complexes (33).

In this study, we propose a new method to estimate the free binding energy between a ligand and receptor. We extend a central idea developed in previous investigations (10, 11) that uses simple descriptors to represent biomolecules as input examples to train a support vector machine (19) and the application of the trained system to previously unseen pairs, estimating their propensity for interaction. Here we seek to learn the function that maps features of a receptor-ligand pair onto their equilibrium free binding energy.

EXPERIMENTAL PROCEDURES
Thermodynamics of Binding

For our purposes, consider that a single protein P binds a single small molecule ligand L to form complex C, or

$$P + L \rightleftarrows C \qquad \text{(Eq. 1)}$$

Assuming that this reaction is in thermodynamic equilibrium, the Gibbs free energy change on binding ΔG^0 is written

$$\Delta G^0 = -RT \ln(K_a) \text{ (J/mol)} \qquad \text{(Eq. 2)}$$

where R is the gas constant, T is the temperature (K), and K_a is the equilibrium binding constant between protein and ligand.[1] K_a is defined as

$$K_a = [C]/([P][L]) \text{ (M}^{-1}) \qquad \text{(Eq. 3)}$$

where $[C]$, $[P]$, and $[L]$ are molar concentrations of complex product, protein, and ligand reactants, respectively. Often the equilibrium dissociation constant K_d is used to quantify ligand binding strength. It is simply the inverse of the binding constant, or

$$K_d = \frac{1}{K_a} = [P][L]/[C] \text{ (M)} \qquad \text{(Eq. 4)}$$

and represents the concentration of ligand required to saturate half of the available binding sites of the protein.

Calculation of ΔG^0 usually entails its partitioning into various energetic components accounting for rotatable bond entropy, hydrogen bonds and ionic interaction forces, lipophilic protein-ligand contact surface, and others (13).

Database of Ligand-Receptor Objects

The data set used in this investigation was aggregated automatically using information located in a number of disparate on-line re-

sources coupled with local computations. An object database was constructed from this data and subsequently sampled to generate examples for training and testing the performance of the regression estimation system. The experimental database consisted of 2,956 objects, each having attributes as summarized in this section.

Ligand-Receptor Complex—Ligand-receptor data were extracted from the Computed Ligand Binding Energy (CLiBE)[2,3] database, a compendium of information on complexed receptors and ligands. Each record in CLiBE contains computed values for the total ligand-receptor potential energy field ΔG^0, given by

$$\Delta G^0 = \Delta G_v + \Delta G_h + \Delta G_e + \Delta G_s \qquad \text{(Eq. 5)}$$

where the right-hand-side partitioning represents energy contributions due to non-bonded van der Waals interactions, hydrogen bonds, electrostatic forces, and ligand desolvation energies, respectively (34). Methods underlying the computation of binding energies comprising the database subject to this investigation are described in Ref. 2.

The complexes within this resource are themselves based on "heterogen" records found in the Protein Data Bank (PDB) (16)[4] for which a chemical identity has been assigned to the ligand. PDB is a public domain repository of experimentally determined structures of biological macromolecules.

Ligand Structures and Chemical Names—Data files with entries representing ligand structures and their associated chemical names were obtained from the NCI, National Institutes of Health Open Database of Compounds.[5] The data entries were represented as "SMILES" strings, where SMILES (Simplified Molecular Input Line Entry System) is a specification and nomenclature for describing molecules as a compact, one-dimensional strings of characters, including atoms, bonds, aromatic rings, and branches (17).

Molecular Connectivity—The SMILES representation for each ligand molecule was converted to a two-dimensional connectivity matrix using a computational chemistry package (JOElib,[6] Ref. 18). The rows and columns of this matrix reflect the cardinality of constituent atoms established by the SMILES representation. At row i and column j, a unit-valued entry is made if the corresponding atoms in the molecule are covalently connected; otherwise the value of that matrix element is zero. Diagonal elements of this matrix store the appropriate atomic number as suggested previously (15).

Molecular Synonyms—To maximize the chemical diversity of objects potentially available for numerical experiments, a list of common chemical synonyms corresponding to each ligand were obtained using the on-line ChemFinder service.[7] Each ligand synonym within its list was used in a lexical similarity search of the NCI, National Institutes of Health compound files to obtain SMILES representations in cases where different chemical names were used for identical ligands across databases.

---

[1] Under physiological conditions (310 K, 1 atm, 1.0 M), the value of RT is about 2.577 kJ/mol or 0.616 kcal/mol.

[2] The abbreviations used are: CLiBE, Computed Ligand Binding Energy; SMILES, Simplified Molecular Input Line Entry System; PDB, Protein Data Bank; nmse, normalized mean square error; nmae, normalized mean absolute error; SVR, support vector regression; HIV, human immunodeficiency virus.

[3] CLiBE circa August 2002 has 14,731 records with 2,803 distinct ligands and 2,256 distinct receptors; see xin.cz3.nus.edu.sg/group/clibe/clibe.asp.

[4] The PDB contains 18,294 structures as of July 23, 2002; see www.rcsb.org/pdb.

[5] Available at cactvs.cit.nih.gov/ncidb2/download.html; this resource currently contains over 250,000 compounds.

[6] Open source available at joelib.sourceforge.net.

[7] See chemfinder.cambridgesoft.com/result.asp.

Support Vector Regression

The support vector algorithm, based on statistical learning theory, is applicable to both 1) binary classification and 2) regression estimation (19). In previous work, we developed methods to train a support vector machine classifier to learn to predict protein-protein interactions using descriptors based on physicochemical properties of paired amino acid sequences (10, 11). In the present application, we propose to exploit the support vector algorithm to solve a regression problem. The concept to be learned is the functional mapping between a set of ligand-receptor features and the total free binding energy of the complex. The basic idea in support vector regression (SVR) is to map a set of input patterns $X = \{x_1, x_2, \ldots, x_l\} \in R^n$ onto a high-dimensional feature space F via a nonlinear mapping $\Phi:R^n \rightarrow R^D$ ($D \gg n$), and then perform linear regression in F. Each pattern vector x_i has a matching target value $y_i \in R$. The goal is to find a function $y = f(x)$ representing the real-valued pairs $\{z_i \mid z_i = (x_i, y_i), i \in 1, \ldots, l\}$ within a certain acceptable maximum deviation level ϵ (12). Practical implementation issues with SVR are presented in Refs. 12 and 14, and theory and algorithms for extension to regression estimation with noisy data appear in Ref. 20.

Feature Representation

Each ligand-receptor complex was transformed into a vector of numerical features presumed salient for learning the target concept. Receptor and ligand feature vectors constructed as outlined in this section are concatenated and labeled with the value of their total free binding energy. These vectors are subjected to support vector machine regression training and cross-validation testing to evaluate how keenly the system learned the concept as posed.

Receptor—Receptor protein features were generated as described previously (10) considering tabulated physicochemical properties (charge, hydrophobicity, and surface tension) of the amino acid sequence thought to be prototypical of binding characteristics of the receptor. Each residue in sequence was replaced by floating point numbers with values corresponding to these physical properties. This vector of numbers was then mapped onto a fixed-length interval to provide a basis for comparison between receptor proteins of varying sequence length.

Ligand—Exemplars for the ligand component of each molecular complex required a novel approach. The design ethos followed here dictates beginning with a minimal, elemental group of features to develop intuition regarding the feature space. In accordance with this approach, the two-dimensional molecular connection matrix described under "Database of Ligand-Receptor Objects" was supplemented by additional arrays, each of which contained numerical values for fundamental, measurable chemical properties characterizing the atoms comprising the molecule. These properties included the atomic *ionization potential energy*, which represents the energy necessary to remove the outermost electron from the ground state of a neutral atom, and the *electron affinity*, which is a measure of energy change upon adding an electron to a neutral atom (21). Ionization energies are always positively valued, while electron affinities may assume either positive or negative numerical values.

For each small molecule ligand, three two-dimensional arrays representing molecular topology, electronic structure, and chemical behavior of the component elements were concatenated into a single, wide matrix. The resulting aggregate data matrix was then factorized using the singular value decomposition (22). The singular values computed in this factorization are extracted, representing a projection onto one-dimensional space of the essential characteristics of molecular bond topology and, it is hypothesized, the spatial distribution of molecular properties important for binding with a receptor.

Burden (15) introduced the idea of computing the eigenvalues of a hydrogen-suppressed molecular bond graph with atomic number on the diagonal and numbers indicating bond presence and type at off diagonal positions. This matrix was used as a means to group substructures for chemical similarity search. In that work, it was maintained that the smallest eigenvalue embodied information on *all* molecules and therefore was sufficient as a topological descriptor. Here all singular values are retained, regardless of their relative magnitudes, as discarding the entire set is not justifiable. This vector is finally stretched (or compressed) onto a fixed-length interval as was performed for the receptor features.

IMPLEMENTATION

Learning Concept—The concept to be learned is the function $y = f(x)$ that maps ligand-protein feature vectors x to the corresponding free energy of binding y. How well the SVR machine learns this concept will be quantified using the statistics described under "Evaluation of Machine Learning" collected from observations of the cross-validation protocol as described under "Cross-validation Experiments."

Evaluation of Machine Learning—One measure of effectiveness for regression estimation is the normalized mean squared error (nmse), given by

$$\text{nmse} = \frac{1}{\sigma^2} \frac{1}{N} \sum_{k=1}^{N} (y_k - \hat{y}_k)^2 \qquad \text{(Eq. 6)}$$

where N is the number of target points predicted, σ^2 is the actual sample variance, and y_k and $\hat{y}_k$ are the actual and estimated target values of the kth data point, respectively (23). Because nmse is normalized by the sample variance, it may be used to compare different regression studies on a more equitable basis than would be possible using the conventional root mean square error; intuitively, a given prediction experiment is less challenging where the variance in the data is small. Notice that if we replace the prediction terms $\hat{y}_k$ with the arithmetic mean $\bar{y}$ in Equation 6, the value of the statistic is 1. This trivial case results when the predictor simply outputs the mean value of the data. Low values of nmse indicate good overall predictive acuity.

Pointwise predictions of ligand binding may be evaluated using the normalized mean absolute error (nmae), defined by

$$\text{nmae} = \frac{1}{\sigma} \frac{1}{N} \sum_{k=1}^{N} \left| y_k - \hat{y}_k \right| \qquad \text{(Eq. 7)}$$

This statistic is normalized by the sample variance for the same reasons as were cited for nmse above. Furthermore, its value may be interpreted as the number of standard deviations, on average, that predictions differ from the target values across the test set. The lower the value of nmae, the better the system pointwise predictive ability.

In some ligand screening situations (such as virtual screening, Ref. 3), predicting the relative ranking of binding strengths among a set of ligand-receptor pairs may be desired. The output of such an analysis would be a list of predicted binding

energies sorted according to predicted magnitudes $\Delta\hat{G}^0$. In such cases a measurement of nonparametric or rank correlation, such as represented by Kendall's τ coefficient (24), is informative. In cross-validation, given an ordered array of N "(actual, predicted)" values $(y_1, \hat{y}_1), \ldots, (y_N, \hat{y}_N)$, we systematically compare the numerical signs of individual bivariate pairs $X = (y_i, \hat{y}_i)$ and $Y = (y_j, \hat{y}_j)$ for $i = 1, \ldots, N, j = (i + 1), \ldots, N$.

If either (a) $y_i > y_j$ and $\hat{y}_i > \hat{y}_j$ or (b) $y_i < y_j$ and $\hat{y}_i < \hat{y}_j$ is observed, X and Y are said to be "concordant." Otherwise the points are "discordant." Kendall's τ expresses the tendency of two ordered lists y and $\hat{y}$ to coordinately increase or decrease and is computed as

$$\tau = \frac{N_C - N_D}{\sqrt{N_C + N_D + T_X} \sqrt{N_C + N_D + T_Y}}, \quad -1 \leq \tau \leq +1 \quad \text{(Eq. 8)}$$

where N_C is the total number of concordant pairs, N_D is the number of discordant pairs, and T_X, T_Y are counts of the "ties" found in X and Y pairs, respectively. A large positive (negative) value of τ indicates that the rank ordered values within $\{y\}$ and $\{\hat{y}\}$ are positively (negatively) correlated.

Cross-validation Experiments—To estimate the generalization error of the trained support vector regression system, we averaged the results of 10 separate 10-fold cross-validation experiments. In k-fold cross-validation, k random, equal-sized, disjoint partitions (folds) of the example data are constructed, and an "inducer" (here, an SVR engine) is trained on $(k - 1)$ folds with the excluded fold being used to test the trained system performance. After k such experiments, the results are averaged, and the observed error rate may be taken as an estimate of the error rate expected upon generalization to new data (25). To reduce further the effects of chance in randomly sampling the data, we averaged the results of 10 different 10-fold cross-validation experiments, performing 100 different training/testing procedures. The results we present are cross-validation averages for the statistics nmse, nmae, and τ as described under "Evaluation of Machine Learning."

The total sample used in these experiments comprised 2,671 distinct ligand-receptor complexes. The output of the trained system is a predicted level of binding free energy y (in kcal/mol) given a set of features abstracted from a given input complex x. A qualitative glimpse of typical results from one complete 10-fold cross-validation test is offered in Fig. 1, which shows a scatter plot of actual *versus* predicted binding energy. Fig. 1 shows that some degree of correlation between prediction and truth exists. This correlation will be examined on an objective basis in the discussion of "Cross-validation Results."

CROSS-VALIDATION RESULTS

The principal results obtained in this investigation are summarized in Table I and in Fig. 1. Table I compares the 10
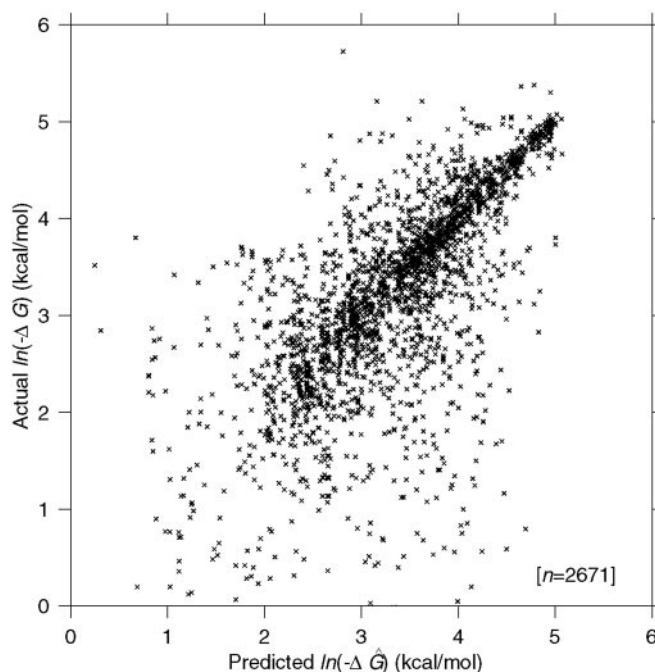

FIG. 1. **Actual *versus* predicted binding free energy.** Shown are typical results from one complete 10-fold cross-validation experiment on the ligand-receptor database discussed under "Database of Ligand-Receptor Objects." Sample size, $n = 2,671$.

10-fold cross-validation error estimates with a number of studies reported in the literature. In contrast to the present results (shown in boldface), all of the competing methodologies shown in Table I are derived from scoring functions or simulations predicated upon knowledge of the three-dimensional structure of receptor and ligand complex. The columns in Table I comprise test sample size N, the mean target binding energy $\bar{y}$, and standard deviation σ^2 in kcal/mol, nmse (Equation 6), nmae (Equation 7), and Kendall's τ (Equation 8).

The records in the table are listed in order of increasing nmse. This statistic is proposed as the primary objective indicator of accuracy for direct prediction of binding free energy.

DISCUSSION

Of particular note on consideration of Table I are the sample size and mean free binding energies characterizing the ligand-receptor data used here when contrasted to the other investigations. The current sample size ($n = 2,671$) is a factor 42 times larger than the next largest data set. The mean free binding energy is seen to be -38 kcal/mol, significantly stronger than the other data summarized in the Table I. Moreover, it can be seen that the present data set is highly variable as the standard deviation (35 kcal/mol) is on the same order as the mean.

Recall from the previous discussion that nmse values on the order of 1 are tantamount to trivial prediction of the mean value of a test data set. Lower values of nmse are associated

Comparison of predictions of ligand-receptor binding free energies in the present investigation (boldface font) and various studies reported in the literature

Test data statistics are sample size N, target mean value $\bar{y}$, and standard deviation σ_y. Results are shown for nmse (Equation 6), nmae (Equation 7), and Kendall's τ (Equation 8). Note: results for the present investigation are average values from 10 10-fold cross-validation experiments.

Source[a]	N	$\bar{y}$	σ_y	nmse	nmae	τ
		kcal/mol	kcal/mol			
1	14	−4.09	1.179	0.198	0.344	0.753
2	12	−0.98	0.332	0.271	0.401	0.667
3	11	−4.25	0.711	0.377	0.466	0.455
4	**2671**	**−37.76**	**35.106**	**0.419**	**0.377**	**0.552**
5	13	−3.93	0.796	0.440	0.497	0.632
6	30	−8.897	2.591	0.720	0.661	0.418
7	17	−8.17	3.785	0.789	0.621	0.358
8	63	−1.45	0.560	1.342	0.836	0.307
9	13	−10.27	6.683	1.466	0.511	0.533

[a] Source numbers refer to the following references: 1, Head *et al.* 1996, Table 3 (26); 2, Böhm 1998, Table 3 (8); 3, Wang *et al.* 1998, Table 4 (4); 4, Bock and Gough 2002 (present investigation); 5, Head *et al.* 1996, Table 4 (26); 6, Wang *et al.* 2002, Table 4 (27); 7, Rarey *et al.* 1996, Table 1 (28); 8, Zhang and Koshland 1996, Table 1 (29); 9, Schapira *et al.* 1999, Table 5 (7).

with genuine learning of underlying patterns in the data and effective generalization. On this basis, the highest predictive accuracy (entry 1, nmse = 0.198) observed in this comparative study was realized by Head and co-workers (26), who present a hybrid approach combining ligand-receptor three-dimensional structural information and parameters derived from molecular mechanics. The test set comprised 14 ligand-receptor complexes.

The second best nmse in this group was achieved by Böhm (8) using a regression-based empirical scoring function based on hydrogen bonds, electrostatics, complementary surface areas, and other characteristics of receptor-ligand pairs where the three-dimensional structure has been previously determined.

Next in our list of prediction results is the investigation reported in Wang *et al.* (4). Their approach uses another empirical scoring function for binding free energy that explicitly accounts for contributions due to Van der Waals interactions, metal-ligand bonding, hydrogen bonds, desolvation energies, and different kinematic effects. A regression equation is developed using these terms derived from known receptor-ligand complexes. All 11 data points in the test sample were based on endothiapepsin receptor complexes.

The current method, based on support vector regression, obtained the fourth best prediction error (nmse = 0.419) averaged over 10 different 10-fold cross-validation tests. We suggest that this error rate represents a significant step for the following reasons.

1. The error rate and rank correlation value are surprisingly competitive with other investigations in light of the relatively large variance and extremely large sample size of the underlying data set. Note that the fifth lowest nmse value in Table I was also obtained by Head *et al.* (26) for a different data set than they used in entry 1.

Group 5 comprised 13 HIV-1 protease-HIV protease inhibitor complexes and showed a value of nmse = 0.440. So the same methodology by the same research group, applied on a different data set, realized much different predictive results. This demonstrates the variability in results that are possible when using small sample sizes while providing confidence in the robustness of our current method and results, which were based on a sample size $n = 2{,}671$.

2. The features used to represent the ligand-protein complexes in the support vector regression do not require any information about three-dimensional structure. All that is required as input data are the amino acid sequence of the receptor, a connection table representing the ligand structure in two dimensions, and the atom characteristics at the nodes of this connection table.

3. There is no limitation on the protein family membership of the putative receptor(s), on the type (organic or synthetic), or on the size of ligand used.

4. The results obtained in this study suggest that it may be possible to infer binding energies for complexes involving newly sequenced or difficult-to-crystallize proteins or for ligands that only exist in computer memory, awaiting synthesis upon successful *in silico* screening.

Rank Correlation—We draw the reader's attention to the trend in Kendall's rank correlation statistic τ in Table I. It is apparent that there is a general inverse correlation between the magnitude of binding energy prediction errors (nmse and nmae) and the value of τ. That is, low values of prediction error are associated with high values of the correlation coefficient. τ measures the tendency of two ordinal random variables

(here actual and predicted binding energy rank) to increase or decrease coordinately. If direct prediction of the physical binding energy is reasonably accurate, we would expect to see a positive and non-trivial correlation between the corresponding rank-ordered variables.

Computing biomolecular binding energies to higher accuracy remains a challenging problem (6). One author recently noted that current computational docking simulations, used to search for the best (lowest energy) "fit" of ligand into a target receptor cavity, still "suffer from insufficient precision of the scoring functions" (5). In Ref. 9, molecular dynamics simulations focused on biotin binding to avidin and streptavidin indicated that the energies of protein and ligand reorganization were found to be significant contributors to protein-ligand binding free energy in molecular dynamic simulations. These reorganization energies were estimated to be on the order of 10–30 and 4.5–6 kcal/mol for protein and ligand, respectively. Because of the large variance in protein reorganization energy, the authors concluded that precise predictions of binding free energy were suspect.

Given these difficulties, the ability to reliably rank a set of ligand-receptor complexes during lead optimization (*versus* directly computing binding energy) remains important in the area of drug discovery. Such a procedure may add value, for example, as a decision aid when down-selecting a set of ligands for chemical synthesis. In connection with the current methodology, we recognize that training the SVR requires example data representing estimated or measured values of binding free energy. The output of a computational technique cannot exceed the accuracy of its input; this is especially true with systems that learn from examples. Therefore, at present the qualitative analysis or ranking of potential ligands may be the main utility of the SVR technique.

The prediction evaluation statistics appearing in Table I are presented in the form of a bar chart in Fig. 2. The investigations numbered along the horizontal axis appear in order of increasing nmse and correspond to the numbering in Table I. This visualization provides a different perspective on the opposing trends of nmse, nmae, and τ as discussed above.

Conclusions—In this work, we have introduced a new methodology, showing that it is possible to predict the binding free energy between ligand and receptor without direct information about their three-dimensional structures.

In cross-validation experiments, we have demonstrated that objective measurements of prediction error rate and rank-ordering statistics are competitive with several other investigations, most of which depend on three-dimensional structural data. The size of the sample used ($n = 2,671$) indicates that this approach is robust and may have widespread applicability beyond restricted families of receptor types. Newly sequenced proteins, or those for which three-dimensional crystal structures are not easily obtained, can be rapidly analyzed for their binding potential against a library of ligands using this methodology.
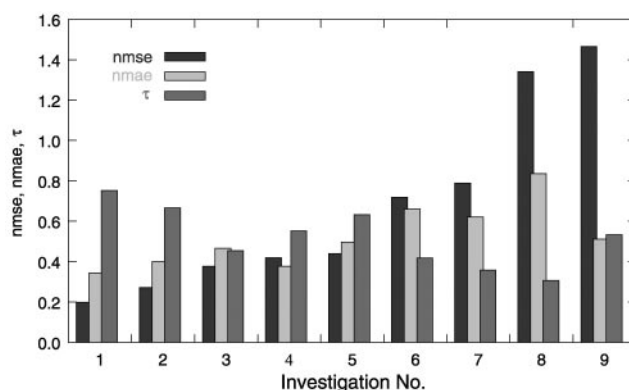

FIG. 2. **Comparison of error and rank correlation statistics between this study and the literature.** The investigations *numbered* along the *horizontal axis* appear in order of increasing normalized mean square error nmse (Equation 6) and correspond to the numbering appearing in Table I. Notice the general trend of inverse correlation between binding energy prediction errors (nmse and nmae) and rank correlation (τ). The present cross-validation results are represented as *Investigation no. 4*.

REFERENCES

1. Augen, J. (2002) The evolving role of information technology in the drug discovery process. *Drug Discov. Today* **7,** 315–323
2. Chen, Y. Z., and Zhi, D. G. (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins Struct. Funct. Genet.* **43,** 217–226
3. Bissantz, C., Folkers, G., and Rognan, D. (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **43,** 4759–4767
4. Wang, R., Liu, L., Lai, L., and Tang, Y. (1998) SCORE: a new empirical method for estimating the binding affinity of a protein-ligand complex. *J. Mol. Model.* **4,** 379–394
5. Kubinyi, H. (2002) The design of combinatorial libraries. *Drug Discov. Today* **7,** 503–504
6. Gillies, M. B. (2001) *Computational Studies of Protein-Ligand Molecular Recognition.* Ph.D. thesis, Universiteit Utrecht, Utrecht, The Netherlands
7. Schapira, M., Totrov, M., and Abagyan, R. (1999) Prediction of the binding energy for small molecules, peptides and proteins. *J. Mol. Recognit.* **12,** 177–190
8. Böhm, H. J. (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from *de novo* design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **12,** 309–323
9. Lazaridis, T., Masumov, A., and Gandolfo, F. (2002) Contributions to the binding free energy of ligands to avidin and streptavidin. *Proteins Struct. Funct. Genet.* **47,** 194–208
10. Bock, J. R., and Gough, D. A. (2001) Predicting protein-protein interactions from primary structure. *Bioinformatics* **17,** 455–460
11. Bock, J. R., and Gough, D. A. (2002) Whole-proteome interaction mining. *Bioinformatics*, in press
12. Smola, A. J., and Schölkopf, B. (1998) *A Tutorial on Support Vector Regression, NeuroCOLT Technical Report NC-TR-98-030*, Royal Holloway College, University of London, UK
13. Böhm, H. J. (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known

three-dimensional structure. *J. Comput.-Aided Mol. Des.* **8,** 243–256

14. Müller, K.-R., Smola, A. J., Rätsch, G., Schölkpof, B., Kohlmorgen, J., and Vapnik, V. (1999) in *Advances in Kernel Methods* (Schölkopf, B., Burges, C., and Smola, A., eds) pp. 243–253, MIT Press, Cambridge, MA

15. Burden, F. R. (1989) Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **29,** 225–227

16. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242

17. Weininger, D. (1988) SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28,** 31–36

18. Wegner, J., and Zell, A. (2002) JOELib: a Java based computational chemistry package, in *6th Darmstädter Molecular-Modelling Workshop*, Technische Universität, Darmstadt, Germany

19. Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*, Springer-Verlag, Heidelberg, Germany

20. Smola, A. J. (1996) *Regression Estimation with Support Vector Learning Machines*. M.Sc. thesis, Technische Universität München, Munich, Germany

21. Boikess, R. S, and Edelson, E. (1981) *Chemical Principles*, 2nd Ed., Harper & Row, New York

22. Golub, G. H., and van Loan, C. F. (1989) *Matrix Computations*, 2nd Ed., Johns Hopkins University Press, Baltimore, MD

23. Gershenfeld, N. A., and Weigend, A. S. (1993) *The Future of Time Series: Learning and Understanding*, Vol. XV, pp. 1–70, Addison-Wesley, Reading, MA

24. Kendall, M. G. (1938) A new measure of rank correlation. *Biometrika* **30,** 81–93

25. Kohavi, R. (1995) in *International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Quebec, August 20–25, 1995*, pp. 1137–1145, Morgan Kaufmann, San Mateo, CA

26. Head, R. D., Smythe, M. L., Oprea, T. I., Waller, C. L., Green, S. M., and Marshall, G. R. (1996) VALIDATE: a new method for the receptor-based prediction of binding affinities of novel ligands. *J. Am. Chem. Soc.* **118,** 3959–3969

27. Wang, R., Lai, L., and Wang, S. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **16,** 11–26

28. Rarey, M., Kramer, B., Bernd, C., and Lengauer, T. (1996) in *Biocomputing: Proceedings of the 1996 Pacific Symposium, Singapore, January 3–6, 1996* (Hunter, L., and Klein, T., eds) World Scientific Publishing, London

29. Zhang, T., and Koshland, D. E. (1996) Computational method for relative binding energies of enzyme-substrate complexes. *Protein Sci.* **5,** 348–356

30. Cunningham, M. J. (2000) Genomics and proteomics: the new millennium of drug discovery and development. *J. Pharmacol. Toxicol. Methods* **44,** 291–300

31. Gohlke, H., and Klebe, G. (2001) Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol.* **11,** 231–235

32. Langer, T., and Hoffmann, R. D. (2001) Virtual screening: an effective tool for lead structure discovery? *Curr. Pharm. Des.* **7,** 509–527

33. Moret, E. E., van Wijk, M. C., Kostense, A. S., and Gillies, M. B. (1999) Scoring peptide(mimetic)-protein interactions. *Med. Chem. Res.* **9,** 604–620

34. Ortiz, A. R., Pisabarro, M. T., Gago, F., and Wade, R. C. (1995) Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **38,** 2681–2691

35. Waszkowycz, B. (2002) Structure-based approaches to drug design and virtual screening. *Curr. Opin. Drug Discov. Dev.* **5,** 407–413