# Nuclear Receptor−DNA Binding Specificity: A COMBINE and Free−Wilson QSAR Analysis[||]

Sanja Tomic,[†,‡] Lennart Nilsson,[§] and Rebecca C. Wade*[,†]

*European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany, Ruder Boskovic Institute, Bijenicka 54, P.O. Box 1016, HR-10001 Zagreb, Croatia, and Center for Structural Biochemistry, Karolinska Institute, NOVUM, S-141 57 Huddinge, Sweden*

*Received July 7, 1999*

Specific binding of transcription factors to DNA is crucial for gene regulation. We derived models for the binding specificity of transcription factors of the nuclear receptor family to DNA using two QSAR methods: a Free−Wilson-like method and COMparative BINding Energy (COMBINE) analysis. The analysis is based on experimental data for the interaction of 20 mutant glucocorticoid receptor DNA-binding domains with 16 different response elements in a total of 320 complexes (Zilliacus, J.; Wright, A. P.; Carlstedt-Duke, J.; Nilsson, L.; Gustafsson, J. A. *Proteins* **1995**, *21*, 57−67). The predictive abilities of the models obtained by the two methods are similar. The COMBINE analysis indicates that the most important properties for determining binding specificity for this dataset are the changes upon binding of the solvation free energies of the bases that are mutated in the dataset and the electrostatic interactions of the mutated nucleotides with certain charged amino acids. Further important descriptors are the changes of solvation free energy and surface area of the side chain of the mutated residue. It is clear, however, that there are additional features important for the specificity of binding that are not included in the models, such as differences in interfacial hydration of the complexes.

## Introduction

The glucocorticoid (GR) and estrogen (ER) receptors are receptors for steroid hormones. They are ligand-inducible transcription factors and members of the nuclear receptor family which also includes receptors for thyroid hormones, retinoic acid, and vitamin D3.[1,2] These proteins have a common amino acid sequence organization composed of discrete functional domains: a highly conserved DNA binding domain (DBD) composed of two zinc fingers and consisting of about 70 residues and a domain for transcriptional regulation. The receptors mainly bind as dimers to partially palindromic repeats of DNA.[3] It is known that the DNA binding specificity of nuclear receptors is mediated, at least in part, by the three variable residues of a region called the P box which is situated at the end of the first zinc finger[3] (Figure 1a).

Specific binding of the DBD of the transcription factor to DNA is crucial for gene regulation. Clearly, the ability to predict binding specificity would be of great value. The factors that govern DNA−protein recognition are complex and incompletely understood.[4] Only recently have sufficient structural data become available to begin to derive, by statistical analysis, quantitative parameters for predicting binding for a set of DNA−protein partners.[5] Most previous analysis has been restricted to computing energy differences for the binding of a small number of protein or DNA sequences differing by single-point mutations. This has been done using molecular dynamics simulation to sample conformation flexibility and free energy perturbation theory for computing relative free energies of binding (see refs 6, 7). Another approach that has been applied to protein−DNA complexes is to compute electrostatic interaction energies by numerical solution of the Poisson−Boltzmann equation and combine these with nonelectrostatic energy terms to estimate binding free energies (see refs 8−12). These approaches are computationally intensive, and approaches amenable to the study of large datasets are necessary.

An alternative route to estimating binding free energies is to use not just structural information but also information about experimentally measured activities or binding affinities for a series of related compounds. The use of this information both demands (because more complexes must be studied) and enables less computationally intensive approaches to be employed. Usually QSARs are derived by correlating information about ligand properties, which may be three-dimensional information, as in the CoMFA (comparative molecular field analysis) method,[13] and activity or binding data (using partial least squares (PLS), neural networks, or similarity indices). More recently, information about the structure of the protein receptor has been used in QSAR

[*] To whom correspondence should be addressed. Tel: +49 6221 387 553. Fax: +49 6221 387 517. E-mail: wade@embl-heidelberg.de.

[†] European Molecular Biology Laboratory.

[‡] Ruder Boskovic Institute.

[§] Karolinska Institute.

[||] Abbreviations: DBD, DNA-binding domain; RE, response element; GR, glucocorticoid receptor; ER, estrogen receptor; GRE, glucocorticoid receptor response element; ERE, estrogen receptor response element; LV, latent variable; PC, principal component; FFD, fractional factorial design. For components of the COMBINE models: SC, side chain; BB, backbone; B, base; S, sugar; P, phosphate; *e*[ ], electrostatic interaction energy term; *v*[ ], van der Waals interaction energy term; $\Delta\Delta G^{hyd}_j$, change on binding of the free energy of solvation of the side chain of residue *j*; $\Delta\Delta G^{hyd\prime}_k$, change on binding of the free energy of solvation of a mutated base *k*; $\Delta SA'_i$, change on binding of the solvent-accessible surface area of nucleotide *i*; $\Delta SA^{np}_j$, change on binding of the nonpolar solvent-accessible surface area of the side chain of residue *j*; $\Delta SA^{p}_j$, change on binding of the polar solvent-accessible surface area of the side chain of residue *j*; $\Delta S^{rot}_T$, change on binding of the total side chain rotational entropy of the DBD; $\Delta S^{rot}_j$, change on binding of the rotational entropy of the side chain of residue *j*.

10.1021/jm9911175 CCC: $19.00 © 2000 American Chemical Society
Published on Web 04/05/2000

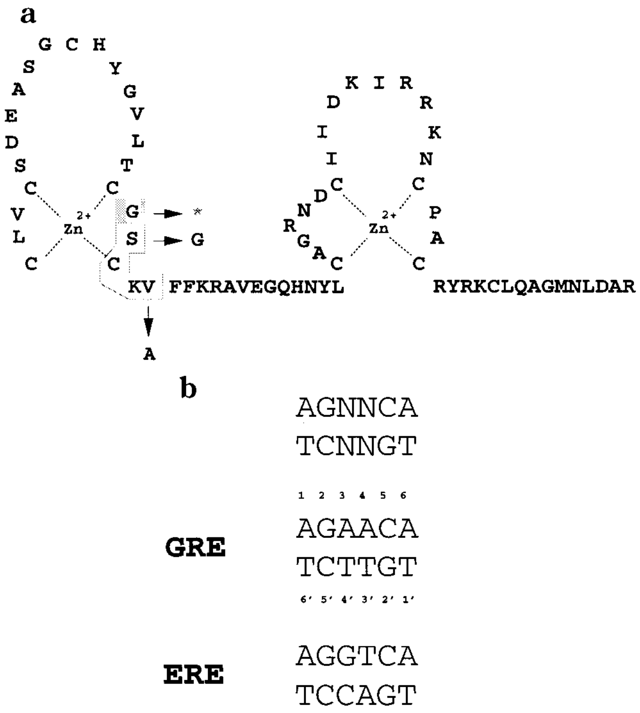

**Figure 1.** (a) Amino acid sequence of the human glucocorticoid receptor DNA binding domain (GR DBD). In the mutated GR DBD for which the specificity of binding was measured, the GR-specific residues in the P box (boxed) Ser-440 and Val-443 are replaced with the ER-specific residues Gly and Ala, and Gly-439 (shaded) is mutated to all possible amino acids. (b) Sequences of the glucocorticoid (GRE) and estrogen (ERE) response elements. The palindromic response element for which specificity of binding of mutated GR DBDs was measured contains all possible base pair combinations (indicated by N in the top sequence) at positions 3 and 4.

studies. In most cases, protein structure information has been used solely for aligning ligands prior to a CoMFA study. In several cases, authors have correlated ligand−protein interaction energies computed by molecular mechanics with binding affinities.[14,15] The binding free energy is, however, rarely a linear function of binding energy. We therefore developed the COMBINE (COMparative BINding Energy) analysis[16,17] approach, in which, instead of correlating the total interaction energy of each complex with binding affinity, the interaction energy is partitioned into fragment- and property-based components. QSARs are derived, using PLS methods,[18] relating binding free energy to a weighted sum of terms dependent on interaction energy components describing physical features important for binding. COMBINE analysis is applied in this work where it is used for the first time for a system consisting of two macromolecules. The macromolecules are mutants of the glucocorticoid receptor and DNA.

Zilliacus et al.[19−22] studied how mutations of a few residues in the GR and ER DBDs modulate DNA binding specificity. The GR DBD binds as a dimer to the glucocorticoid response element (GRE), a partially palindromic repeat consisting of two hexameric half-sites with a three-base-pair spacing inbetween (Figure 1b).[21,23] The hexameric half-sites bound by the GR and ER differ from each other at the two central base pair positions, i.e., at the bases at positions 3 and 4 in the REs (Figure 1b).[24] The initial preferential binding of the GR DBD to one of the two half-sites facilitates binding of the dimer to the other half-site.[24−26] Zilliacus et al.[22] performed measurements on 20 mutants of the GR differing in the amino acid at position 439 which is positioned close to the two central base pairs in the RE in the 3D structure of the GR DBD−DNA complex determined by X-ray crystallography[27] (Figure 2). They measured how this particular mutation modulates binding affinity to 16 different REs, i.e., GREs in which the two central base pairs were mutated to the remaining 15 combinations of the four nucleotides naturally occurring in DNA. We used their functional data from a transactivation assay performed on 320 complexes in Free−Wilson-like QSAR analysis[28,29] and COMBINE analysis.[16,17]

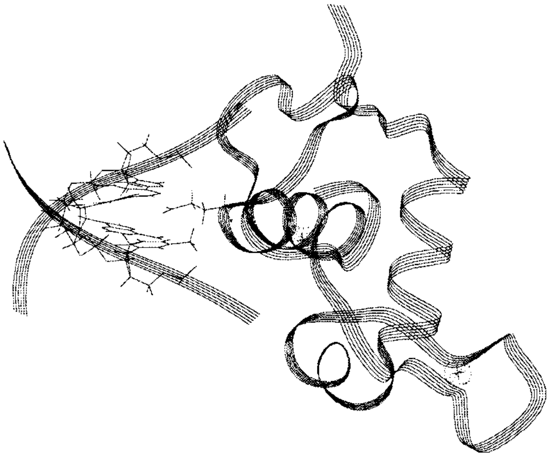

**Figure 2.** Ribbon diagram of a mutant DBD−DNA complex. All atoms of the mutation sites, residue 439 (mutated to Glu) and the central base pairs (TA, CG), are shown.

**Aim of This Study.** The aim of this work is to derive models predictive of the specificity of binding of transcriptor factor DBDs to DNA REs. These models should enable the physicochemical features that are important for binding specificity to be elucidated, thus providing useful insights for the design of DBD−DNA interactions. The models should provide a quantitative link between three-dimensional structural models of the DBD−DNA complexes and the experimental observables of binding and activity. They should thus enable objective interpretation of binding and activity data using three-dimensional structural models.

A Free−Wilson-like analysis was performed previously on a different set of 8 GR mutants binding to four different REs.[19] The model obtained on the set of 32 objects had good predictive ability, although its interpretation was complicated. We performed a similar Free−Wilson-like analysis for the present dataset. This analysis is fast to perform and does not require any models of three-dimensional structures. This analysis provided a baseline and guide for subsequent COMBINE analysis, with which a variety of models, dependent on different properties of the system, were derived. While predictive abilities are similar for the two methods, COMBINE analysis provides physical insights into the factors determining binding specificity.

## Methods

**(0) Experimental Data.** DNA binding affinities were assigned as the logarithmic ($\log_{10}$) values of the activities obtained in an in vivo transactivation assay for 320 DBD−

DNA combinations.[22] A previous study by Zilliacus et al.[19] showed reasonable consistency between directly measured in vitro binding affinities and the logarithmic values of the activities obtained in a transactivation assay for similar DBD–DNA interactions.

**(I) Free–Wilson Analysis.** The Free–Wilson approach was the first QSAR procedure to be developed.[28,29] In this approach, it is assumed that biological activity can be described by the additive properties of substituents on a basic molecular structure. In addition to a model incorporating only additive occupation variables, we also used a Free–Wilson-like approach incorporating association variables that was applied earlier to a set of wild-type and mutant GR DBDs binding to the GRE, ERE, and two intermediate sequences.[19]

Two different models were constructed in which DNA binding affinities were described by: (a) 28 occupation variables (for each of the 20 possible amino acid residues at position 439 of the DBD and each of the 4 possible pairs of nucleotides at positions 3 and 4 of the DNA half-site) and (b) 28 occupation and 480 association variables, making 508 variables in total. The association variables consist of 160 variables describing each combination possible between an individual nucleotide pair (at either position 3 or 4) and the amino acid at position 439 and 320 variables describing each combination possible between both pairs of nucleotides and the amino acid at position 439. A value of 1 or 0 was assigned to each variable, corresponding to the presence or absence, respectively, of a particular residue or base(s) or combination thereof. To determine how the amino acid residue at position 439 of the GR DBD and the base pairs at positions 3 and 4 of the mutated GRE contribute to the DNA binding specificity, i.e., to derive occupation and association coefficients, the values were subjected to PLS analysis[30] as implemented in the GOLPE program.[31,32] Validation of the model was performed by two methods: (a) by cross-validation and (b) by dividing the dataset into a training set and an external test set and testing on the latter.

**(II) COMBINE Analysis.** The basis of COMBINE analysis[16,17] is the assumption that the protein–receptor (in our case DBD–DNA) binding free energy, $\Delta G$, can be approximated by a weighted sum of $n$ selected terms, $\Delta u_i^{\mathrm{sel}}$, each describing the change in property $u_i$ upon binding:

$$\Delta G = \sum_{i=1}^{n} w_i \Delta u_i^{\mathrm{sel}} + C \qquad (1)$$

The weights, $w_i$, are determined by PLS analysis[18] and, if necessary variable selection, as implemented in the GOLPE program.[31,32]

The expression used in this work to provide an estimate of the DBD–DNA binding free energy is:

$$\begin{aligned}\Delta G = {} & \sum_{i=1}^{n\mathrm{DNA}} \sum_{j=1}^{n\mathrm{DBD}} w_{ij}^{\mathrm{vdw}}\, \Delta u_{ij}^{\mathrm{vdw}} + \sum_{i=1}^{n\mathrm{DNA}} \sum_{j=1}^{n\mathrm{DBD}} w_{ij}^{\mathrm{ele}}\, \Delta u_{ij}^{\mathrm{ele}} + \\ & \sum_{j=1}^{N\mathrm{DBD}} w_j^{\mathrm{hyd}}\, \Delta\Delta G_j^{\mathrm{hyd}} + \sum_{k=1}^{4} w_k^{\mathrm{hyd}'}\, \Delta\Delta G_k^{\mathrm{hyd}'} + \sum_{j=1}^{N\mathrm{DBD}} (w_j^{\mathrm{p}}\, \Delta SA_j^{\mathrm{p}} + \\ & w_j^{\mathrm{np}}\, \Delta SA_j^{\mathrm{np}}) + \sum_{i=1}^{N\mathrm{DNA}} w_i^{\mathrm{sa}'}\, \Delta SA'_i + \sum_{j=1}^{N\mathrm{DBD}} w_j^{\mathrm{rot}}\, \Delta S_j^{\mathrm{rot}} + \\ & w^{\mathrm{conf}}\, \Delta E^{\mathrm{conf}} + w^{\mathrm{conf}'}\, \Delta E^{\mathrm{conf}'} + C \qquad (2)\end{aligned}$$

where $\Delta u_{ij}$ is an intermolecular energy term between group $i$ in the DNA and group $j$ in the DBD; $\Delta\Delta G^{\mathrm{hyd}}{}_j$ and $\Delta\Delta G^{\mathrm{hyd}'}{}_k$ are the changes on binding of the free energy of solvation of the side chain of residue $j$ and of the base of the mutated nucleotide $k$, respectively (see eq 3 below); $\Delta SA_j{}^{\mathrm{p}}$ and $\Delta SA_j{}^{\mathrm{np}}$ are the changes on binding of the polar and nonpolar, respectively, solvent-accessible surface areas of the amino acid residue $j$; $\Delta SA'_i$ is the change on binding of the solvent-accessible surface area of the nucleotide $i$; $\Delta S^{\mathrm{rot}}{}_j$ is the change on binding of the rotational entropy of residue $j$; $\Delta E^{\mathrm{conf}}$ and $\Delta E^{\mathrm{conf}'}$ are the changes in potential energy of the protein and DNA, respectively, upon binding; $n$DNA and $n$DBD are the number of groups in the DNA and DBD, respectively; and $N$DNA and $N$DBD are the number of residues and nucleotides in the DBD and DNA hexameric half-site, respectively. For this study, groups are defined as follows. Each amino acid residue except the N-terminal one was split into two groups: side chain and backbone; each nucleotide was split into three groups: base, phosphate, and sugar. So the following intermolecular interaction energies were considered for each residue and nucleotide pair: side chain–base, backbone–base, side chain–phosphate, backbone–phosphate, side chain–sugar, and backbone–sugar.

Not all terms were used in all models. The terms $\Delta E^{\mathrm{conf}}$ and $\Delta E^{\mathrm{conf}'}$, describing the conformational changes of the molecules on binding, were omitted from most models as they are poorly correlated ($R < 0.1$) with measured activities and do not improve predictive ability. Single conformations of each DNA, DBD, and DNA–DBD complex were modeled with molecular mechanics force fields and it was assumed that these are representative of the full ensemble of structures that would be sampled in solution. However, to treat the conformational changes of molecules upon binding more thoroughly, a more detailed modeling of bound and free molecules should be performed than was possible here. The considerable size of the system and number of complexes to be modeled prevented a detailed characterization of conformational changes in this case. Eriksson and Nilsson have performed simulations of the GR DBD and two of its mutants[33] and of the ER DBD,[34] free in solution and bound to the RE, and have found that 500–1000 ps of simulation can be required to observe conformational changes in the DBDs.

**(a) Building Models of the DNA–DBD Complexes.** Systems consisting of the 73-residue long mutated GR DBD (Figure 1a) bound to 6-base pair (bp) mutants of the GRE were modeled[22] (Figure 1b). The structures of the mutated complexes were derived from the crystallographically determined structure of the rat GR DBD–DNA complex (Luisi et al., 1991). The sequence of the rat GR DBD is identical to that of the human GR DBD used in the transactivation assays,[22] the results of which are utilized in this work.

The mutants were modeled as follows. Residues Ser-440 and Val-443 in the specifically bound protein monomer (DBD1) from the crystal structure were replaced with the corresponding side chains from the ER, Gly-440 and Ala-443. The experiments were performed on mutants of this ER–GR hybrid DBD because its binding specificity is not restricted to the GRE. This made it possible to examine the influence of any amino acid introduced at position 439 on specificity of binding to different response elements. Gly-439 of this DBD (Figure 1a) was replaced with the other 19 naturally occurring amino acids, and the base pairs at the central positions (3 and 4, Figure 1b) of the GRE were mutated to a total of 16 different combinations. In this way, 320 structures, one for each possible residue/base combination, were obtained.

**(b) Energy Minimization of the DNA–DBD Complexes.** Initially, all complexes were energy minimized for 290 steps with restraints with the CHARMM program;[35] details of the procedure are described elsewhere.[22] The complexes as well as each macromolecule on its own were further energy minimized using the AMBER 4.0 program with the all-atom force field.[36] Minimization was performed without constraints using a distant-dependent dielectric constant. The parameters for the $\mathrm{Zn}^{2+}$ ion (partial atomic charge: +2e; Lennard–Jones $\sigma$ and $\epsilon$ parameters: 1.22 Å and 0.25 kcal/mol; Zn–S bond length and force constant: 2.229 Å and 50 kcal/mol/Å$^2$) were derived from the results of MOPAC 6.0 AM1[37,38] calculations using the standard procedure available in the AMBER 4.0 program. The coordinates of the DBDs, REs, and their complexes, obtained after 30, 70, and 300 steps of energy minimization with the AMBER program, were utilized for the COMBINE analysis. More complete minimization was not carried out as artifacts might be introduced because the minimizations are done in the absence of explicit water molecules and for a protein monomer rather than the dimer for which experiments were performed.

**(c) Estimation of the Free Energy of Hydration.** It is known[39] that the change of entropy due to release of water molecules and ions as well as the change of conformational entropy upon binding influence binding affinity. To estimate the change of solvation free energy on binding, the following procedure was adopted. For the side chains of amino acid residues, the change in solvation free energy was calculated using the modified Fauchere−Pliska[40] values determined by Wimley et al.[41] For mutated bases, values calculated by Kollman et al.[42] and recently determined experimentally by Shih et al.[43] were used. The computed and experimentally determined values of the relative solvation free energies differ, but the trends are the same and the models obtained with each of the sets of values are similar.

The following expression was used to calculate the free energy change:

$$\Delta\Delta G_i^{\mathrm{hyd}} = \Delta G_i^{\mathrm{hyd}}(\Delta SA_i/SA_i^{\mathrm{T}}) \qquad (3)$$

for an amino acid side chain or mutated base, $i$. $\Delta G_i^{\mathrm{hyd}}$ is the relative free energy of solvation of the whole side chain/base, and $SA_i^{\mathrm{T}}$ is its total solvent-accessible surface area. For amino acids, values of $\Delta G_i^{\mathrm{hyd}}$ relative to Gly in the pentapeptide AcGG-X$_i$-GG [40] were used. For bases, values of $\Delta G_i^{\mathrm{hyd}'}$ relative to thymine[42] were used. For the latter, $SA_i^{\mathrm{T}}$ is the solvent-accessible surface area of the base in an isolated nucleotide calculated with the NACCESS program.[44,45] The change of the solvent-accessible surface area per residue/base upon binding ($\Delta SA_i$) was also calculated using the NACCESS program.

Additionally, the change upon binding of the solvent-accessible surface area of each nucleotide from the palindromic unit of the REs and the change of polar and nonpolar surface areas of each amino acid side chain ($\Delta SA^{\mathrm{np}}$, $\Delta SA^{\mathrm{p}}$) were determined and included in the model as separate $X$-variables.

**(d) Estimation of Side Chain Rotational Entropy.** The change upon binding of the conformational entropy, $\Delta S^{\mathrm{rot}}$, of the protein was approximated by the change in side chain rotational entropy upon binding DNA. This was estimated by sampling torsional degrees of freedom in the unbound and bound DBD. The Biopolymer module of the InsightII program package[46] was used for this purpose. For each residue $j$:

$$\Delta S_j^{\mathrm{rot}} = \ln \frac{nR_j^{\mathrm{co}}}{nR_j^{\mathrm{p}}} \qquad (4)$$

where $nR_j^{\mathrm{co}}$ and $nR_j^{\mathrm{p}}$ are the number of accessible rotamers of the residue $j$ in a complex and in the free protein, respectively. The cutoff energy for considering a rotamer as accessible was 1000 kcal/mol. Other cutoff energies were tested, but the results were similar.

**(e) Data Preparation for PLS Analysis with the GOLPE Program.** The matrix of $X$-variables contained 320 rows (1 for each complex) and 10541 columns assigned as follows: 1−5148, van der Waals interaction energies between the groups ($\Delta u_{ij}$); 5149−10296, electrostatic interaction energies between the groups ($\Delta u_{ij}$); 10297−10367, $\Delta\Delta G_j^{\mathrm{hyd}}$ per amino acid residue $j$; 10368−10371, $\Delta\Delta G_k^{\mathrm{hyd}'}$ per mutated base $k$; 10372−10517, $\Delta SA_j^{\mathrm{p}}$ and $\Delta SA_j^{\mathrm{np}}$ for each residue $j$; 10518−10537, $\Delta SA_i'$ per nucleotide $i$ (for the mutated nucleotides, the base, sugar, and phosphate groups were considered separately); 10538−10539, $\Delta S^{\mathrm{rot}}{}_{439}$ (the change of rotational entropy of residue 439) and $\Delta S^{\mathrm{rot}}{}_{\mathrm{T}}$; 10540−10541, $\Delta E^{\mathrm{conf}}$ and $\Delta E^{\mathrm{conf}'}$.

Models were constructed using either the whole set of $X$-variables described above or subsets of it. When both inter- and intramolecular interactions were considered, a different matrix of $X$-variables containing 32105 columns was used, which included the columns listed above and additional columns to describe intramolecular energy terms.

The 1D **Y**-matrix was the same as for the Free−Wilson analysis and consisted of 320 logarithmic values of activities from in vivo transactivation assays.[22]

**(f) Chemometric Analysis of the Data.** (i) Before PLS analysis, the data were pretreated using different procedures: zeroing ($X$-variables, i.e., columns in the **X**-matrix, with values below a certain threshold were eliminated), applying a minimum standard deviation cutoff (columns in the **X**-matrix with a standard deviation below a certain threshold were eliminated), and block scaling using block unscaled weights[32] (the **X**-matrix was divided into blocks of related data, with blocks scaled either so that the same total weight was given to each or so that differences in weights between blocks of variables were reduced) or autoscaling (the **X**-matrix was transformed so that each column of data had an average of 0 and a standard deviation of 1).

(ii) PLS[18] statistical analysis was performed to derive models and determine the variables that are the most important for specificity of binding.

(iii) The models were subjected to internal and external validation. For internal validation, cross-validation using 5 random groups and 20 randomizations as implemented in the GOLPE program[32] was used. The predictive ability of a model was characterized by SDEP and $Q^2$ values:

$$SDEP = \sqrt{\sum \frac{(Y - Y')^2}{N}} \qquad (5)$$

$$Q^2 = 1 - \frac{(Y - Y)^2}{\sum (Y - \langle Y \rangle)^2} \qquad (6)$$

where $Y$ is experimental activity, $Y'$ is predicted activity, $\langle Y \rangle$ is the average experimental activity, and $N$ is the number of DBD−DNA complexes. External validation was performed for test sets of objects. In addition, the predictive ability of the models was checked in computations with $Y$-values scrambled and with $Y$-values assigned random numbers.

(iv) To extract the most predictive $X$-variables, variable selection was performed for some models by fractional factorial design (FFD) as implemented in the GOLPE program; 2−3 cycles of FFD were performed according to the improvement in $Q^2$ computed by the cross-validation procedure described above. Dummy variables were introduced with the ratio of true variables to dummies set to 4:1 and a design combinations-to-variables ratio of 2:1. Variables which were determined to be noise were excluded and uncertain variables were retained.

**(g) Datasets Used in the Analysis.** Besides the models obtained for the entire set of 320 objects (complexes), models were also derived for reduced data sets. The 289- and 41-object datasets were selected based on the results of the Free−Wilson analysis (see Results section). Further datasets were selected according to the chemical properties of the mutated residue in the DBD.

## Results

**(I) Free−Wilson Analysis. (a) 320-Object Dataset.** Free−Wilson QSAR analysis was performed in order to determine how the amino acid residue at position 439 of the GR and the base pairs at positions 3 and 4 of the GRE mutants contribute to DNA binding specificity. The results are summarized in Table 1.

The $Y$-variable is the logarithmic value of the level of transactivation observed in yeast cells expressing DBDs. In the model, the mean $Y$-value is normalized to 0. Thus, a positive coefficient indicates that a particular $X$-variable improves binding, while a negative coefficient indicates the negative influence of the $X$-variable on binding. The $X$-variables that mainly determine the first latent variable (LV), which is the only significant component of the model, are the nucleotide occupancy variables: the presence of T and A at positions 4′ and 3′, respectively, of the RE improves binding, while the presence of G and C at any of these positions decreases

**Table 1.** Predictive Performance of the Free−Wilson-like Models

| $N^a$ | $Y^b$ | $X^c$ | LV$^d$ | SDEP | $Q^2$ | SDEC | $R^2$ |
|---|---|---|---|---|---|---|---|
| 320 | L | o + a (508) | 2 | 0.58 | 0.42 | 0.43 | 0.68 |
| 320 | L | o + a (508→313)$^e$ | 3 | 0.53 | 0.51 | 0.50 | 0.75 |
| 320 | S | o + a (508) | 1 | 0.67 | 0.22 | 0.62 | 0.33 |
| 320 | R | o + a (508) | 1 | 0.45 | −0.20 | 0.35 | 0.26 |
| 320 | L | o (28) | 2 | 0.59 | 0.41 | 0.54 | 0.50 |
| 289 | L | o + a (508) | 1 | 0.51 | 0.54 | 0.46 | 0.62 |
| 289 | L | o + a (477→408)$^e$ | 1 | 0.50 | 0.55 | 0.46 | 0.61 |
| 289 | S | o + a (508) | 1 | 0.67 | 0.22 | 0.62 | 0.33 |
| 289 | R | o + a (508) | 1 | 0.47 | −0.25 | 0.38 | 0.17 |
| 289 | L | o (28) | 1 | 0.51 | 0.54 | 0.49 | 0.57 |
| 41 | L | o + a (508) | 3 | 0.24 | 0.86 | 0.04 | 0.99 |
| 41 | L | o + a (134→130)$^e$ | 3 | 0.24 | 0.87 | 0.04 | 0.99 |
| 41 | S | o + a (508) | 1 | 0.68 | −0.05 | 0.37 | 0.65 |
| 41 | R | o + a (508) | 1 | 0.41 | −0.27 | 0.24 | 0.57 |
| 41 | L | o (28) | 3 | 0.15 | 0.95 | 0.08 | 0.98 |

[a] Number of complexes. [b] Activity: L = logarithmic (base 10) value of the measured activity, S = scrambled logarithmic activities, R = random numbers. [c] Free-Wilson-like *X*-variables: o = occupation (28), a = association (480). The number of *X*-variables is given in parentheses. [d] Number of LVs giving highest value of $Q^2$. Note that for models derived with the 320-object dataset only the first LV is significant, and for the 41-object dataset only two LVs are significant. For both datasets, adding one further LV results in an increase of only ca. 0.02 in the value of $Q^2$. [e] Model obtained after FFD variable selection with the number of selected variables shown after the arrow.

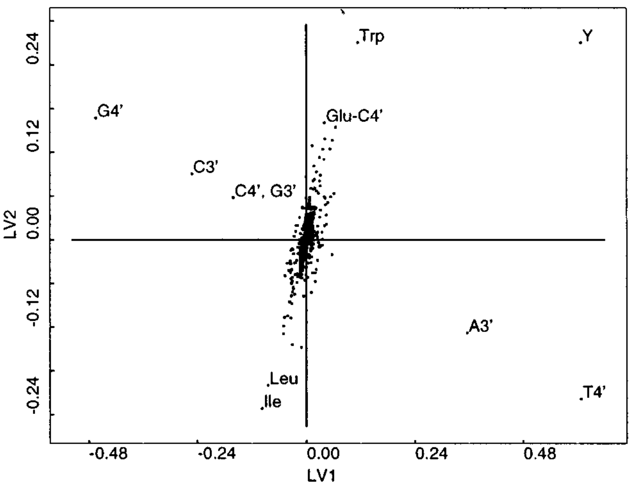

**Figure 3.** Partial weight plot for the *X*-variables and the activity (*Y*) in the first two LVs obtained for the entire dataset by Free−Wilson QSAR analysis for the model listed first in Table 1.

binding (Figure 3). The second LV is mainly determined by the type of amino acid at position 439 of the DBD. The presence of Trp improves binding, while the presence of Ile or Leu decreases binding. This is in accord with the attractive interactions between Trp and G and C at positions 3 and 4 noted earlier by Zilliacus et al.[22] It is also consistent with the observation that removal of the two highest activity objects, Trp-GT and Trp-GC, from the dataset reduced the contribution of the second LV and the highest $Q^2$ value (0.40) was obtained with one LV. The same is valid after further removal of the next highest activity object, Tyr-GT. The most important association variable is Glu-C4′ which positively defines the second LV. This interaction is present in the ER−ERE complex. Its identification is also in agreement with the specific and attractive hydrogen-bond interaction found between a Glu side chain at residue 439 and a C base at nucleotide 4′ when using the COMBINE procedure (see below).

Although after FFD *X*-variable selection, the number of *X*-variables was reduced to 313, the distribution of *X*-variables in the partial weight plot of the first and second LVs did not change significantly. The main difference is that the Glu-C4′ association variable becomes less important in the partial weight plots of these LVs. However, the number of LVs for which the predictive ability of the model reaches a maximum increased after variable selection from two to three, and the third LV is mostly determined by the Glu-C4′ variable (positively) and by the Glu-T4′ variable (negatively).

The Free−Wilson model derived for the 320-object dataset with only the 28 occupancy variables fits the data less well and is slightly less predictive than the models derived with both occupancy and association variables.

**(b) Reduced Datasets.** While the models derived with the 320-object dataset are robust to scrambling of *Y*-values or using random *Y*-values, predictive ability is limited. However, after removing the 31 objects with the lowest predictive ability from the original set of data, leaving 289 objects, the internally cross-validated $Q^2$ values of all models improved (e.g. for 508 *X*-variables, $Q^2$ changed from 0.42 to 0.54; see Table 1). Similar improvement is noticed in the COMBINE model when switching from the entire dataset to this reduced dataset (see below). The majority of the objects excluded from the reduced dataset have extreme values of activity, mostly low, suggesting that they are likely to provide less reliable data points.

By further reduction of the number of objects, i.e., by selecting only the 41 objects for which very good predictive ability is obtained with the Free−Wilson model, it was possible to obtain a model with excellent fit and very high internal predictive ability that is robust to scrambling of activities or using random activities (e.g. for 508 *X*-variables, $Q^2 = 0.86$, SDEP = 0.24; see Table 1). When this reduced dataset was then used in COMBINE analysis, models with similar internal predictive ability were obtained (see Table 4 below). The predictive ability of this Free−Wilson model was tested on four datasets: the entire dataset, the 289-object dataset, and external datasets with 279 and 248 objects. The latter two datasets are complements of the 41-object dataset to the entire dataset and to the 289-object dataset, respectively. The respective SDEP values determined for these datasets with the model derived for the 41-object dataset using 508 variables are 0.59, 0.50, 0.63, and 0.53. These values are similar to the internal SDEP values determined for the 320- and 289-object datasets. Similar results are obtained using COMBINE analysis (see below). Thus the 41-object dataset appears to represent the full dataset well and with less “noise”. It contains most of the DBDs (16 out of 20), and a majority of the response elements (13 out of 16) are present. Missing are complexes with Arg, Asp, Glu, and Leu and those with response element nucleotides 3′ and 4′ being CG, AG, and TC.

After FFD *X*-variable selection on models derived with the reduced datasets, the number of LVs for which the maximum internal predictive ability of the model is obtained remains unchanged, and the internal and external predictive abilities of the models change little.

The dominant *X*-variables in the models derived for the reduced datasets are, as for the models for the entire dataset, nucleotide occupation at positions 3′ and 4′ with A and T positively and C and G negatively defining the first LV. This finding is in accord with the results obtained by Plaxco and Goddard.[47] They found that any protein−DNA complex that excludes solvent from the base pairs will select AT (or TA) because of shielding of the thymine methyl group from solvent. However, the second LV in the model obtained for entire dataset and the second and third LVs in the model obtained for the 289-object dataset are defined by the amino acid occupation variables as well.

**(II) COMBINE Analysis. (a) Sensitivity of COMBINE Models to Implementation Parameters.** To derive a reliable 3D-QSAR model, different atomic coordinates, types of pretreatment, variable selection, and datasets were evaluated.

**(i) Influence of structure optimization on the predictive ability:** To check the influence of structural optimization on the models, the complexes (objects) optimized by 30, 70, and 300 energy minimization steps with the AMBER 4.0 program were analyzed. The quality and the structure of the QSAR models without intramolecular interaction energy terms are similar for the datasets with 320 objects optimized with 30 and 70 steps. The models obtained for the dataset with objects optimized with 300 steps have a slightly different structure, and their fitting and predictive abilities are lower. However, when both inter- and intramolecular interactions were used, models with similar predictive ability were obtained for all three datasets. Overall, the models were best with the coordinates optimized with the 70-step energy minimization, and these were used for subsequent models.

**(ii) Influence of different variable pretreatments on the predictive ability:** Pretreatment was performed to reduce noise by zeroing and imposing a minimum standard deviation cutoff; i.e., the *X*-variables with low absolute values ($\leq 0.1$) and standard deviations ($\leq 0.05$) were excluded from the PLS analysis. This reduced the number of *X*-variables (from ca. 10000 to ca. 300) so that the model could be subjected to the FFD variable selection procedure and the degree of uncertainty introduced by the modeling procedure could be minimized. This pretreatment had no significant influence on the predictive ability and structure of the models. Indeed, the *X*-variables selected as important in the COMBINE models have absolute values and variances that are at least several times the cutoff thresholds; thus variables of small magnitude or variance are not overemphasized in the models.

Autoscaling does not improve predictive performance in any model. Application of block unscaled weights results in models for all datasets with similar internal predictive performance but fewer latent variables than models derived without scaling. A range of block scaling parameters was tested. Both default scaling factors between blocks and more conservative choices of scaling factors were tested, but this choice rarely resulted in significant changes in external predictive measures or the structure of the models, and the models with the best external predictive ability were chosen. The greatest scaling factor was applied to the free energy of solvation block: scaling factors of 3−10 (default value) times the factor for the block of intermolecular energy terms were used. Block scaling with electrostatic and van der Waals interactions treated as separate blocks of *X*-variables did not improve the predictive ability of models. The results of the PLS analysis performed separately on van der Waals and electrostatic interactions show that the predictive ability of the former interactions is lower than of the latter (see Table 2). However, block scaling with the van der Waals and electrostatic interactions as one block has a positive effect on the predictive performance of the models. The relative importance of different kinds of *X*-variables in these models is better balanced and physically more justified when block scaling is performed. The predictive ability of these models is higher for a few LVs than when block scaling is not used and is robust as tested by scrambling of the activity data, as illustrated for the ISA model in Table 2.

**(iii) Variable selection:** Variable selection by the FFD strategy was performed for selected models, and it improved their performance. For most of the models derived from all datasets, it resulted in reduction of the optimal number of LVs (see, e.g., the IG model in Table 2). For models derived for the reduced datasets (289 and 41 objects), variable selection also increases both internal and external predictive performance; see, e.g., Figure 4. For the 41-object dataset, variable selection has a greater positive effect on models excluding intramolecular terms than models including them.

**(iv) Datasets considered:** Three datasets were considered in detail: the full 320-object dataset and the reduced datasets of 289 and 41 objects. The objects in the reduced datasets were selected on the basis of the results of the Free−Wilson QSAR analysis and are the same as in the reduced datasets in the Free−Wilson analysis. Further datasets were selected on the basis of the chemical properties of the mutated residue in the DBD.

The main structure of the partial weight plots for the models derived for the 289-object dataset is very similar to that determined for the entire dataset. For the models obtained for the 41-object dataset, similarity exists but the parameters related to residues other than that mutated are more important in these models. Models obtained with the 41-object dataset were tested on the entire dataset and the 289-object dataset as well as external datasets which are complements of the 41-object dataset to these: the 279- and 248-object datasets. The external SDEPs for these datasets are similar or slightly better than those determined internally, using the same models (Table 4). This is in accord with the observation that the main structure of the models, i.e., the part of the model described with only the first few LVs (up to 5), is similar for all datasets considered. Note that the exclusion of the highest activity objects does not influence the robustness of the models: the 289-object dataset lacks the two highest activity objects (Trp-GT and Trp-GC) and the 41-object dataset lacks the three highest activity objects (Trp-GT, Trp-GC, and Tyr-GT). Some models were also derived with a 317-object dataset consisting of the full dataset with the three

**Table 2.** Predictive Performance of the COMBINE Models Derived with the Complete 320-Object Dataset

| model identifier | pretreatment[a] | FFD var sel | $Y$[b] | vdw inter[c] | ele inter[c] | vdw intra[d] | ele intra[d] | $\Delta\Delta G^{\text{hyd}}$[e] | $\Delta SA$[f] | $\Delta S^{\text{rot}}$[g] | LV[h] | SDEP | $Q^2$ | SDEC | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [i] | P | | L | + | | | | | | | 4 | 0.73 | 0.08 | 0.71 | 0.14 |
| [i] | P | | L | | + | | | | | | 4 | 0.65 | 0.27 | 0.62 | 0.33 |
| I0 | P | | L | + | + | | | | | | 5 | 0.65 | 0.27 | 0.61 | 0.36 |
| I0 | P | | L | + | + | | | | | | 13 | 0.58 | 0.42 | 0.48 | 0.60 |
| I0 | P | + | L | + | + | | | | | | 5 | 0.60 | 0.38 | 0.56 | 0.47 |
| IG | PB | | L | + | + | | | + | | | 5 | 0.60 | 0.38 | 0.57 | 0.45 |
| IG | PB | | L | + | + | | | + | | | 11 | 0.57 | 0.44 | 0.51 | 0.54 |
| IG | PB | + | L | + | + | | | + | | | 5 | 0.57 | 0.43 | 0.54 | 0.49 |
| ISA | P | | L | + | + | | | | + | | 5 | 0.69 | 0.21 | 0.63 | 0.30 |
| ISA | PB | | L | + | + | | | | + | | 5 | 0.66 | 0.26 | 0.62 | 0.35 |
| ISA | P | | L | + | + | | | | + | | 13 | 0.56 | 0.45 | 0.50 | 0.57 |
| ISA | PB | | L | + | + | | | | + | | 13 | 0.56 | 0.46 | 0.49 | 0.59 |
| ISA | PB | | S | + | + | | | | + | | 7 | 0.71 | 0.13 | 0.67 | 0.24 |
| IGSA | PB | | L | + | + | | | + | + | | 5 | 0.59 | 0.40 | 0.55 | 0.47 |
| IGSA | PB | | L | + | + | | | + | + | | 10 | 0.56 | 0.46 | 0.50 | 0.55 |
| IGSA | PB | + | L | + | + | | | + | + | | 4 | 0.57 | 0.44 | 0.54 | 0.50 |
| IGSAS$_{\text{rot}}$ | PB | | L | + | + | | | + | + | + | 4 | 0.60 | 0.38 | 0.56 | 0.45 |
| IGSAS$_{\text{rot}}$ | PB | + | L | + | + | | | + | + | + | 4 | 0.59 | 0.40 | 0.56 | 0.46 |
| II0 | P | | L | + | + | + | + | | | | 5 | 0.59 | 0.40 | 0.55 | 0.48 |
| IIG | PB | | L | + | + | + | + | + | | | 7 | 0.57 | 0.44 | 0.52 | 0.53 |
| IIGSA | PB | + | L | + | + | + | + | + | + | | 5 | 0.59 | 0.41 | 0.53 | 0.51 |

[a] Types of $X$-variable pretreatment: P, zeroing (0.1) and minimum standard deviation cutoff (0.05) (for detailed explanation, see Methods section); B, block scaling of $X$-variables with the van der Waals and electrostatic interactions considered as one block. [b] Activity: L = logarithmic and S = scrambled values of measured activity. [c] Intermolecular van der Waals and electrostatic interactions. [d] Intramolecular van der Waals and electrostatic interactions. [e] Changes of the free energy of solvation (per amino acid residue and mutated base). [f] Changes of the surface area (per amino acid residue and nucleotide). [g] Changes of the rotational entropy of the mutated amino acid residue and the DBD. [h] Number of LVs; when the SDEP continues to steadily decrease as the number of LVs exceeds 8, values are given for the model with the lowest determined SDEP and the model with 5 LVs. [i] Coordinates after a 70-step energy minimization was used for all models except those marked, for which coordinates after 30 steps of energy minimization were used.

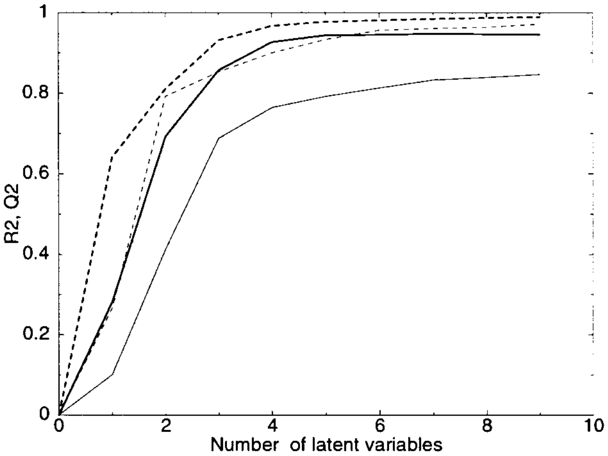

**Figure 4.** Development of predictive measures during derivation of the IGSA model as a function of the number of LVs. Key: fitted $R^2$, dashed line; internal cross-validated $Q^2$, solid line; variable selection, before − thin line, after − thick line.

highest activity objects removed. The performance and structure of the models did not change significantly from that of models for the full dataset.

**(b) COMBINE Models Derived Without Intramolecular Energy Terms as *X*-Variables.** The best COMBINE model, which we name the IGSA model for the $X$-variables contributing to it (see Tables 2−4), was obtained by using three blocks of $X$-variables: the partitioned van der Waals and electrostatic intermolecular interactions, the change in the free energy of solvation, and the change in solvent-accessible surface area terms. Addition of terms describing the change of side chain rotational entropy per residue did not improve predictive ability (see Tables 2−4), and after variable selection, side chain rotational entropy terms did not appear as important $X$-variables. We first describe the I0 models which have only intermolecular interaction energy terms, as do models in other applications of the COMBINE method,[16,17,48,49] and thus provide reference models. Then we describe the effect of adding other $X$-variables to the models.

**(i) I0 models (intermolecular energy terms only):** Only models derived without data scaling are discussed, as separate block scaling of the partitioned electrostatic and van der Waals intermolecular energy terms decreases predictive ability (see above). Without variable selection, the I0 models have low predictive ability when fewer than 6 LVs are considered, but their predictive ability increases steadily with the number of LVs (Table 2), indicating that the specificity of binding of transcription factors to DNA is a complex process.

To determine the structure of the $X$-variables and the distribution of objects based on it, we performed principal component (PC) analysis. In the score plot of the first and second PCs, the complexes with Arg and Lys at residue 439 are distinguished from those with Glu and Asp in the first PC and the other complexes are clustered inbetween (Figure 5). Among the complexes with Glu and Asp at residue 439, those with a DNA RE with cytosine at position 4′ are separated, primarily in the second PC, from all the others. This is probably because a hydrogen bond is formed between the cytosine base and the carboxylic group of the residue. In the score plot of the third and fourth PCs, the complexes are evenly distributed and there is no clustering in the middle of the plot (not shown).

The main PCs and the most important LVs are described with the same $X$-variables. In all I0 models, the first two LVs and PCs are defined mainly by the following interactions: electrostatic and van der Waals SC439−B4′ (side chain of residue 439 in DBD-base of

**Table 3.** Predictive Performance of the COMBINE Models for the 289-Object Dataset

| model identifier | pretreatment[a] | FFD var sel | $Y^b$ | inter[c] | intra[d] | $\Delta\Delta G^{\text{hyd }e}$ | $\Delta SA^f$ | $\Delta S^{\text{rot }g}$ | LV[h] | SDEP | $Q^2$ | SDEC | $R^2$ | SDEP[i] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I0 | P | | L | + | | | | | 7 | 0.57 | 0.42 | 0.52 | 0.52 | |
| I0 | P | + | L | + | | | | | 5 | 0.52 | 0.52 | 0.48 | 0.58 | 0.57 |
| IG | PB | | L | + | | + | | | 7 | 0.53 | 0.50 | 0.49 | 0.56 | |
| IG | PB | + | L | + | | + | | | 4 | 0.52 | 0.52 | 0.50 | 0.57 | 0.56 |
| ISA | PB | | L | + | | | + | | 7 | 0.54 | 0.47 | 0.49 | 0.57 | |
| IGSA | PB | | L | + | | + | + | | 4 | 0.53 | 0.51 | 0.49 | 0.56 | |
| IGSA | PB | + | L | + | | + | + | | 5 | 0.50 | 0.55 | 0.46 | 0.61 | 0.56 |
| IGSAS$_{\text{rot}}$ | PB | + | L | + | | + | + | + | 4 | 0.52 | 0.52 | 0.48 | 0.58 | 0.58 |
| II0 | P | | L | + | + | | | | 4 | 0.51 | 0.53 | 0.48 | 0.59 | |
| IIG | PB | + | L | + | + | + | | | 4 | 0.51 | 0.54 | 0.48 | 0.58 | 0.59 |
| IIGSA | PB | + | L | + | + | + | + | | 3 | 0.52 | 0.52 | 0.49 | 0.57 | 0.55 |

[a–h] See the corresponding footnotes in Table 2. [i] SDEP for the entire 320-object dataset.

**Table 4.** Predictive Performance of COMBINE Models for the 41-Object Dataset

| model identifier | pretreatment[a] | FFD var sel | $Y^b$ | inter[c] | intra[d] | $\Delta\Delta G^{\text{hyd }e}$ | $\Delta SA^f$ | $\Delta S^{\text{rot }g}$ | LV[h] | SDEP | $Q^2$ | SDEC | $R^2$ | SDEP$^{(320,\ 279,\ 289,\ 248)}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I0 | P | | L | + | | | | | 5 | 0.52 | 0.29 | 0.18 | 0.81 | 0.69, 0.74, 0.65, 0.70 |
| I0 | P | | L | + | | | | | 13 | 0.34 | 0.71 | 0.07 | 0.99 | |
| I0 | P | + | L | + | | | | | 4 | 0.28 | 0.81 | 0.15 | 0.95 | 0.60, 0.66, 0.52, 0.59 |
| IG | PB1 | | L | + | | + | | | 7 | 0.29 | 0.78 | 0.17 | 0.93 | 0.66, 0.71, 0.58, 0.63 |
| IG | PB1 | + | L | + | | + | | | 5 | 0.17 | 0.93 | 0.12 | 0.97 | 0.58, 0.62, 0.50, 0.53 |
| ISA | PB1 | | L | + | | | + | | 5 | 0.49 | 0.39 | 0.30 | 0.77 | 0.72, 0.76, 0.68, 0.73 |
| ISA | PB1 | + | L | + | | | + | | 5 | 0.30 | 0.76 | 0.17 | 0.93 | 0.60, 0.64, 0.53, 0.56 |
| IGSA | PB1 | | L | + | | + | + | | 5 | 0.27 | 0.80 | 0.18 | 0.92 | 0.62, 0.66, 0.52, 0.56 |
| IGSA | PB1 | + | L | + | | + | + | | 5 | 0.15 | 0.94 | 0.10 | 0.98 | 0.59, 0.63, 0.50, 0.54 |
| IGSAS$_{\text{rot}}$ | PB1 | + | L | + | | + | + | + | 4 | 0.19 | 0.91 | 0.12 | 0.97 | 0.61, 0.65, 0.53, 0.57 |
| II0 | P | | L | + | + | | | | 5 | 0.13 | 0.96 | 0.08 | 0.98 | 0.61, 0.65, 0.50, 0.54 |
| II0 | P | + | L | + | + | | | | 5 | 0.12 | 0.97 | 0.08 | 0.98 | 0.61, 0.65, 0.50, 0.54 |
| IIG | PB1 | + | L | + | + | + | | | 5 | 0.12 | 0.97 | 0.08 | 0.98 | 0.61, 0.65, 0.50, 0.54 |
| IISA | PB1 | + | L | + | + | | + | | 2 | 0.13 | 0.96 | 0.10 | 0.98 | 0.61, 0.65, 0.51, 0.55 |
| IIGSA | PB1 | + | L | + | + | + | + | | 4 | 0.12 | 0.96 | 0.09 | 0.98 | 0.61, 0.65, 0.50, 0.54 |

[a–i] See corresponding footnotes in Tables 2 and 3. In the last column, SDEP values are given for the 320-, 279-, 289-, and 248-object datasets, respectively. The 279- and 248-object datasets are external datasets.

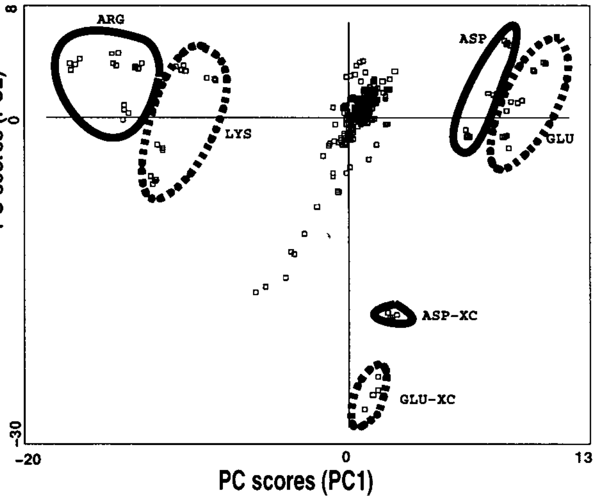

**Figure 5.** Distribution of the complexes in the score plot of the first and second PCs of the I0 model obtained for the entire dataset by COMBINE analysis. Encircled objects are labeled by the identity of residue 439. XC represents a response element with cytosine at the 4′ position.

nucleotide 4 in the upper part of the half-binding site), electrostatic SC439–S4′ (sugar 4′), and van der Waals SC439–B3′ interactions (Figure 6a). With refinement of the structures by 300-step energy minimization with the AMBER program, LV 1 is additionally defined by the electrostatic SC439–S3′, SC470–P2′ (phosphate group of nucleotide 2′), and SC470–P3′ interactions. The electrostatic interactions between the zinc ion and the 4′ sugar and phosphate groups also contribute to LVs 1–4. LVs 3 and 4 are also defined by the electrostatic and van der Waals interactions between SC439 and the sugar and phosphate groups of nucleotides 3′ and 4′ and the electrostatic interactions between the side chain of Lys-442 and the bases of the mutated nucleotides (Figure 6b).

The structure of the partial weight plots for the 289-object dataset is almost identical to that for 320 objects. In contrast, in the partial weight plot of the first two LVs for the 41-object dataset, the electrostatic interaction energy SC447–B3′ appears as important. The same term appears as important in the partial weight plot of LVs 3 and 4 for the most energy-minimized (300 steps) dataset for 320 objects after variable selection. From analysis of the loading plots, it can be seen that the more energetically optimized a dataset is, the more amino acid residues appear as important for the specificity of binding; i.e., their importance for discriminating between different REs increases. The reason for this is that conformational differences between complexes increase with the amount of minimization.

After variable selection for the 320-object dataset (70-step energy minimization), not only terms relating to the mutated components but also other terms, such as for the electrostatic interactions between the $Zn^{2+}$ ion and nucleotide 4′ and between the bases of the mutated residues and the side chains of the residues Glu-427, His-432, Lys-442, Lys-446, Arg-447, Glu450, and Arg-470, appear as important descriptors (Figure 7). Interactions involving residue 439 only become important in LVs 3 and 4. Consistent with these data, in the crystal structure[27] and during MD simulations,[50] Lys-442 and

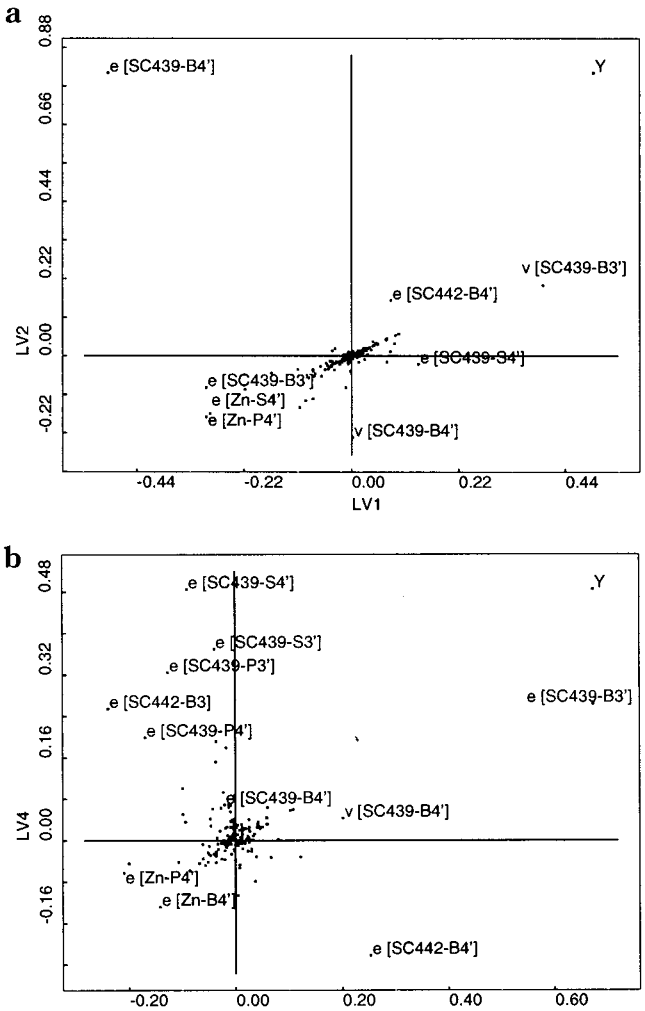

**Figure 6.** Partial weight plot for the *X*-variables and the activity (*Y*) in (a) the first two LVs and (b) the third and fourth LVs of the I0 model obtained for the entire dataset by COMBINE analysis.

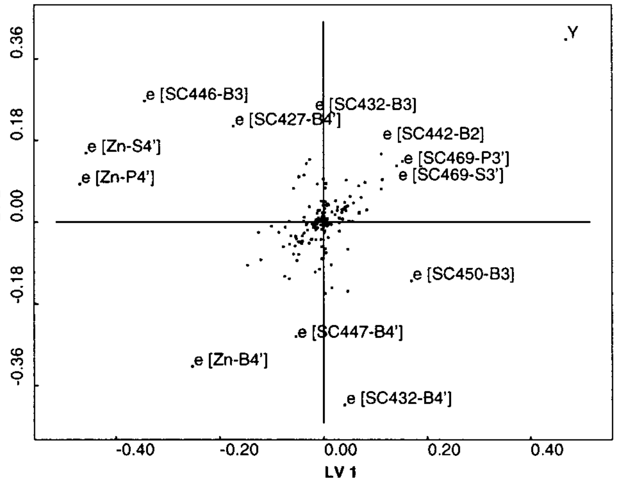

**Figure 7.** Partial weight plot for the *X*-variables and the activity (*Y*) in the first two LVs of the I0 model obtained for the entire dataset by COMBINE analysis after variable selection.

Arg-446 form direct contacts with the REs, and Arg-470 is connected to DNA via water molecules.

**(ii) IG, ISA, and IGSA models (intermolecular interaction energy with free energy of solvation and/or solvent-accessible surface area terms):** The IG, ISA, and IGSA models have better predictive ability than the I0 model for all datasets considered when block scaling is applied (Tables 2−4).

In the IG model for the full dataset, LV 1 is described mainly by the decrease upon binding of $\Delta\Delta G^{\mathrm{hyd}}{}_{3'}$ and $\Delta\Delta G^{\mathrm{hyd}}{}_{4'}$, as expected from the results of the Free−Wilson model. LV 1 is also defined by $\Delta\Delta G^{\mathrm{hyd}}{}_{3}$ in the 41-object dataset model. LVs 3 and 4 are additionally described by $\Delta\Delta G^{\mathrm{hyd}}{}_{4}$, $\Delta\Delta G^{\mathrm{hyd}}{}_{439}$, and the electrostatic SC439−B4′ and SC439−B3′ interaction energies. The negative change of the free energy of solvation of the bases upon binding is correlated with increasing activity in this model. Thus, the score in the first LV results in complexes with AT nucleotide pairs predominantly having the highest activity and those with GC nucleotide pairs having the lowest activity.

In the ISA model, the most important LVs are described by the changes of the solvent-accessible surface areas of the mutated nucleotides and the mutated amino acid residue and the electrostatic interaction energy between the mutated amino acid residue and the mutated base B4′.

The IGSA models with block scaled *X*-variables have better internal and external predictive ability than the IG and the ISA models (see Tables 2−4). The internal predictive ability of the model for the entire 320-object dataset steadily increases with the number of LVs and reaches its maximum (before variable selection) for 10 LVs (SDEP = 0.56, $Q^2$ = 0.46; Table 2). However, the lowest SDEP (0.53, $Q^2$ = 0.51) in the 289-object dataset is obtained with only 4 LVs (Table 3). Thus, as for the Free−Wilson model, the maximum predictive ability of the IGSA model for the 289-object dataset is reached with fewer LVs than for the entire dataset. However, the *X*-variables that have a high influence on the models are very similar for these datasets (Figure 8). The important *X*-variables in the IGSA model are those already mentioned for the I0 and IG models and, in addition, the changes of the solvent-accessible surfaces areas of the mutated bases, the nucleotide 2′, and the mutated amino acid 439.

After FFD variable selection, the internal and external predictive abilities of the models improved (Tables 2 and 4) and the distribution of *X*-variables in the partial weight plots changed. As for the I0 model obtained after variable selection, the importance of the terms (intermolecular interaction energy, hydration free energy, and solvent-accessible surface area) related to the mutated amino acid residue decreases and the importance of the terms related to the charged amino acid residues (Glu-427, Lys-442, Lys-446, Arg-447, Glu-450, Arg-470, and Arg-477) increases (Figure 8). However, the terms related to solvation of the mutated bases remain as important *X*-variables. In the IGSA model, the terms describing the change of the solvent-accessible surface area of the mutated nucleotides, mainly the mutated base pair B4−B3′ and nucleotide 2, are important *X*-variables before and after the FFD variable selection. However, the terms describing the change of the solvent-accessible surface area of the mutated residue ($\Delta SA^{\mathrm{np}}{}_{439}$, $\Delta SA^{\mathrm{p}}{}_{439}$) disappear from the partial weight plots of the IGSA model after variable selection

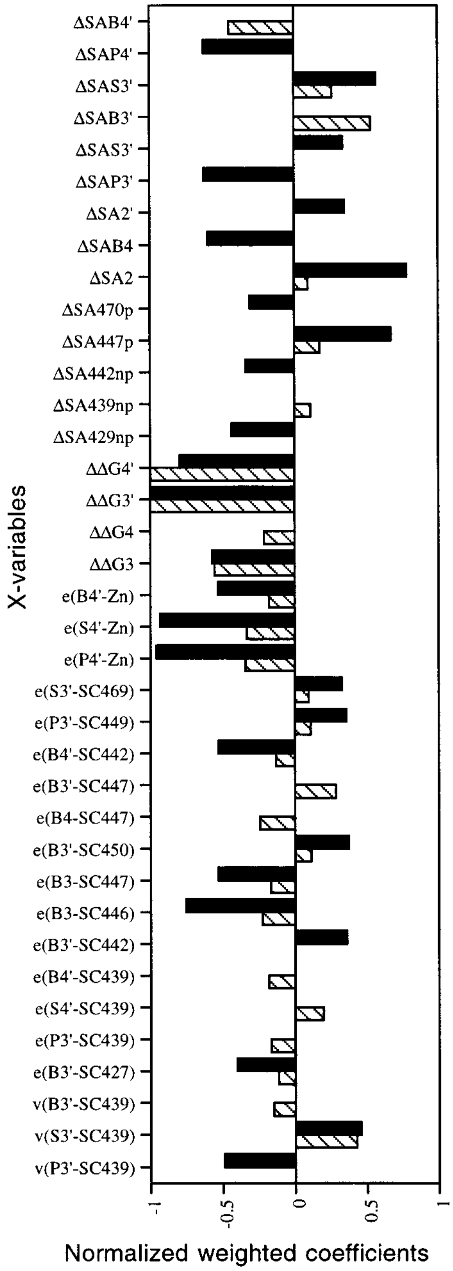

**Figure 8.** Normalized weighted coefficients for the most important *X*-variables in the IGSA model for the 289-object dataset before (hatched) and after (filled) variable selection. Normalization is performed with respect to the largest coefficient for each model.

and the term $\Delta SA^{p}{}_{447}$ appears as an important *X*-variable instead.

Only the models derived with the 41-object datasets could be tested on large external test sets (see Table 4). The correlation coefficient between experimental and predicted activities for the IGSA model derived for the 41-object dataset is 0.75 for the 289-object dataset and 0.65 for the 320-object dataset. For the complement external datasets, the correlation coefficients are 0.73 for the 248-object dataset and 0.62 for the 279-object dataset. Predicted activities are plotted against experimental activities in Figure 9.

**(c) COMBINE Models Including Intramolecular Energy Terms as *X*-Variables.** It was noticed that the conformational energy of the complex as well as of the hexameric DNA binding half-site alone is quite well-correlated with the logarithmic value of the activity ($R \sim 0.5$). However, neither the change in the total conformational energy of the DNA and the protein upon binding nor its components are correlated with the logarithm of the activity ($R < 0.1$). Inclusion of the latter energies into models does not improve their predictive ability (data not shown). In contrast, inclusion of the conformational energy of the complex into models increases their predictive ability. Because of this, we decided to use the partitioned intramolecular energy of the complex in the COMBINE analysis. Physically, this quantity is not expected to be directly related to the free energy of binding, although recent data have shown that the binding energy of a molecule to a protein with a loose structure can derive in part from changes occurring within the protein.[51]

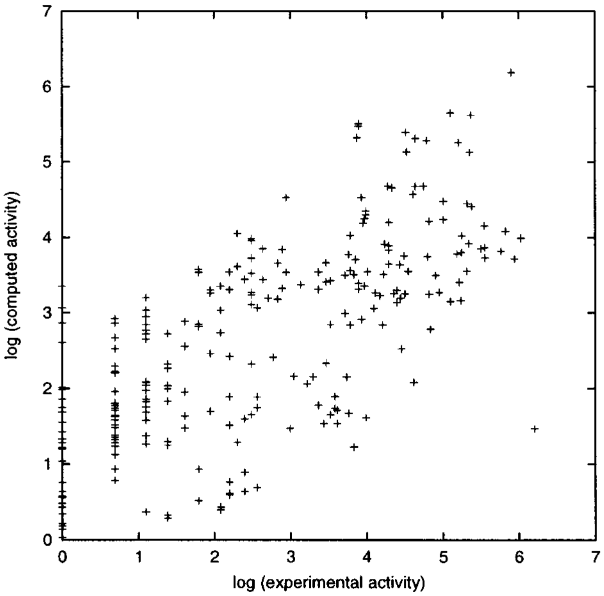

**Figure 9.** Plot of predicted against experimental log activities computed with the IGSA model derived with the 41-object dataset and applied to the external 248-object complement dataset.

**(i) II0 models (inter- and intramolecular interaction energy terms only):** The predictive abilities of the II0 models are, when considering the smaller number of LVs, significantly higher than that of the I0 models (see Tables 2−4). The predictive performance is similar to that of the IG and IGSA models, i.e., models obtained using $\Delta\Delta G^{\mathrm{hyd}}$ terms.

In the II0 model, the most predictive variables are the electrostatic and van der Waals intramolecular interaction energies between the paired bases of the mutated nucleotides. However, it is not the decrease of intramolecular interaction energies but the increase that is correlated with activity. Obviously, the predictive performance of these variables is not directly related to their physical meaning but points to an implicit effect, presumably the hydration free energy of the mutated bases and the conformational change of the DNA. LV 2 of the II0 model is additionally described by the electrostatic intermolecular SC439−B4′ and the intramolecular B4′-S4′ and B3−S3 interaction energies. LVs 3 and 4 are defined by both the DNA and the protein intramolecular interaction energy terms as well as by the electrostatic intermolecular interaction energies between the mutated amino acid residue and other

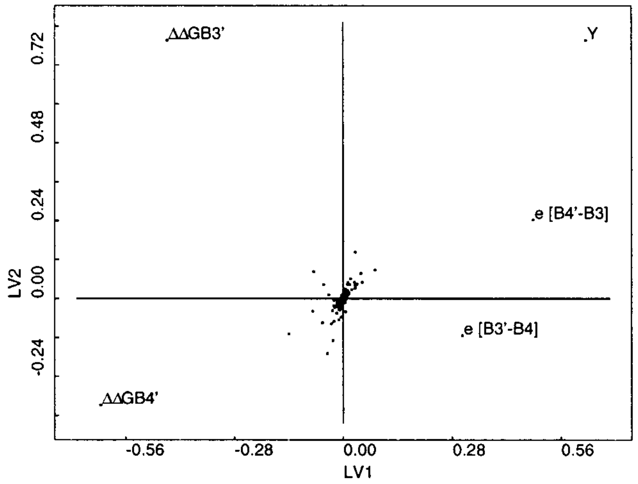

**Figure 10.** Partial weight plot for the *X*-variables and the activity (*Y*) in the first two LVs of the IIG model obtained for the entire dataset.

predominantly charged amino acid residues and the mutated nucleotides.

The score in LV 1 results in complexes in which both mutated nucleotide pairs are AT having high activity and those in which both pairs are CG having low activity. The other complexes are inbetween, but the complexes with T or A at position 4′ are predicted to bind more strongly than those with T or A at position 3′. Thus, the presence of nucleotides A and T in the RE enhances binding. This was also found in the IG and IGSA models and in the Free−Wilson models, although in the II0 and the Free−Wilson models, discrimination of the complexes on the basis of the identity of the mutated nucleotides is more apparent than in the IG and IGSA models.

From inspection of the score plots obtained by PC analysis, it is obvious that the mode of interaction of the DBD and DNA also strongly depends on the identity of the mutated residue. Arg and Lys at the mutated residue have the opposite effect on activity to Asp and Glu, and these charged amino acids are distinguished from other amino acids. Further discrimination between complexes shows that the binding specificity depends largely on the identity of the mutated bases and on intermolecular electrostatic interactions between the DNA and some charged DBD residues.

FFD variable selection does not significantly increase the predictive ability (internal or external) of the II0 model. In addition, the most important *X*-variables, the electrostatic intramolecular energy terms, B3−B4′, B4−B3′, B4′−S4′, and B3−S3, remain unchanged upon variable selection.

**(ii) IIG, IISA, and IIGSA models (with inter- and intramolecular interaction energy terms and combinations of free energy of solvation and/or solvent-accessible surface area terms):** The internal and external predictive abilities of these block scaled models are, for all datasets, similar to that of the II0 model (Tables 2-4). The first LV of the block scaled IIG model is described by the $\Delta\Delta G_{3'}$ and $\Delta\Delta G_{4'}$ values, negatively, and by the electrostatic intramolecular interaction energies B4′−B3 and B3′−B4, positively (see Figure 10). As expected, the negative change in free energy caused by the replacement of an AT (or TA) pair by CG (or GC) is correlated with the positive change of the intramolecular interaction energy between the paired bases. Accordingly, the score plot of LV 1 and activity is the combination of corresponding PLS plots of the IG and the II0 models.

In the IISA model, LV 1 is predominantly determined by the electrostatic B3−B4′ and B4−B3′ interactions and LV 2 is mostly defined by the changes of surface area of the mutated amino acid residue (polar and nonpolar) and nucleotides. In the partial weight plot of LVs 3 and 4, *X*-variables that appear as important are the electrostatic interaction energies between the side chain of the mutated residue and the mutated nucleotides.

FFD variable selection has no significant influence on the predictive performances of all these models. Important *X*-variables besides the electrostatic intramolecular energy terms, B3−B4′, B4−B3′, B4′−S4′, and B3−S3, are terms ($\Delta\Delta G_3$, $\Delta\Delta G_4$, $\Delta SA_{B3}$, $\Delta SA_{B3'}$, $\Delta SA^{np}_{443}$, and $\Delta SA^{np}_{446}$) describing the free energies of hydration and solvent-accessible surface areas of the mutated bases and some charged amino acid residues.

## Discussion

The best COMBINE model for all the datasets analyzed is obtained by using the partitioned intermolecular interaction energies, the free energy of solvation, and the solvent-accessible surface area terms (IGSA model). In this model, the most predictive *X*-variables, obtained after FFD variable selection which significantly increases its internal and external predictive abilities, are the electrostatic energy terms describing the intermolecular interaction between the mutated nucleotides, mostly their bases, and the mutated residue 439, the $Zn^{2+}$ ion, and the side chains of the amino acid residues, Glu-427, Lys-442, Lys-446, Arg-447, Glu-450, Arg-470, and Arg-477, the change of solvation free energies of the mutated bases B3′ and B4′, and the change of the solvent-accessible surface areas of the B4−B3′ base pair and the bases 2 and 5 (Figure 8). The same *X*-variables appear as predictive after variable selection in the COMBINE models that include additional *X*-variables such as intramolecular energy and side chain rotational entropy. Inclusion of the latter did not improve the predictive abilities of the models, and after variable selection, side chain rotational entropy terms disappeared from the partial weight plots of the low LVs. On the other hand, addition of the solvation free energy term to any combination of *X*-variables generally increases the predictive ability of the model.

Comparison of the Free−Wilson and COMBINE models shows a similarity between the former and the IG and II0 models. The first LV in all these models is determined by the mutated nucleotides: in the Free−Wilson model by the occupancy variables of the mutated nucleotides, in the IG model by the change of the free energy of solvation of the mutated bases, and in the II0 model by the electrostatic intramolecular interaction energies between the mutated base pairs. The presence of either A or T at the mutation sites is predicted by all these models to enhance binding. The physical meaning of these variables is most clear in the IG model. Apparently, for the set of transcription factor DBDs considered in this work, the change of the free energy of solvation of the DNA is a dominant influence on the specificity of binding. Eriksson and Nilsson[6] estimated the difference in Gibbs free energy of three different REs in solution and in complex with the GR DBD. They found that the affinity of the GRE decreases after replacement of the T3A4′ base pair with the C3G4′ base pair. This is consistent with more beneficial desolvation of the GRE(T3A4′), and the likelihood that this is due to the thymine methyl group is supported by the study of Plaxco and Goddard.[47] The COMBINE analysis also shows that this is probably due to the hydrophobic thymine methyl group as the partial weight plots of the IIG model show that an increase of the intramolecular electrostatic energy between paired bases is accompanied by a decrease of the free energy of the mutated bases and correlates with an increase in activity. However, the specificity of binding is defined by other physical features as well. In the majority of models obtained before variable selection, the electrostatic interaction between SC439 and B4′ is the dominant interaction energy term. In addition, other electrostatic (as well as some van der Waals) interaction energies between different groups of the mutated nucleotides and the side chain of the mutated residue, a few other charged amino acid residues and the $Zn^{2+}$ ion, often appear as important. The importance of intramolecular terms in the II0 model also hints at the important role of conformational changes of both DNA and the DBD for specificity of binding. This would be consistent with a recent study of transcription factor−DNA binding.[33,34] Note, however, that interactions of the protein backbone rarely appear as important *X*-variables in the models indicating that the protein−DNA recognition mechanism for this system is direct rather than indirect readout.

For the dataset we studied in the present work, binding affinities were not available, so instead we used in vivo transactivation capacities. Although the authors of the previous QSAR analysis[19] found that in vivo transactivation capacity is generally consistent with the level of in vivo DNA binding in their dataset, this might not be true for the entire dataset that we used. Better predictions of the experimental values for the 320- and 289-object datasets are obtained with the models derived for the 289- and 41-object datasets, respectively, than with the model derived on the larger dataset itself. Consistent with this, for 4 models for each physicochemical class of amino acids at residue 439 (neutral, alkyl, positive, negative), models with internal predictive $Q^2$ values up to 0.5 could be obtained for all residue types except “negative” (i.e. Asp and Glu) for which no predictive models could be derived. Moreoever, of the 20 models for each residue type, the worst model, which was not predictive, was for Asp. Complexes containing Asp and Glu at position 439 are absent from the 41-object dataset. While this may point to some systematic errors in structural modeling or energy computation, it is worth noting that most of the complexes omitted in the 289-object dataset had low activities, pointing to the presence of noise which obstructs derivation of a model with good predictive performance for the entire dataset. On the other hand, from comparison of the previous Free−Wilson analysis[19] with our Free−Wilson analysis, it seems that the identity of residue 439 is not sufficient to define the specificity of binding of transcription factors to DNA. It would not be surprising that binding specificity is more completely defined in Free−Wilson analysis with the set of complexes analyzed in ref 19 with DBDs in which three variable positions were considered: 439, 440, and 443, than within the set of complexes with the DBDs varying only at position 439.

Kosztin et al.[52] performed MD simulations of two ER DBD−DNA systems, one consisting of the ER DBD dimer and the consensus segment of DNA and the other including the dimer and the nonconsensus segment of DNA. Comparing these two systems, they found that binding specificity is the result of a network of protein−DNA hydrogen bonds (direct and water-mediated). The protein residues identified in the network are Glu-439, Lys-442, Lys-446, Arg-447, Arg-470, Lys-471, and Arg-477. These residues were also identified as making protein−DNA hydrogen bonds in simulations of the ER DBD−ERE system by Eriksson and Nilsson.[34] Some of these residues, particularly Lys-442, Arg-447, and Arg-470, appear as important variables in our COMBINE models.

In the investigations presented here, we have not included explicit water molecules that, according to MD simulations of the ER DBD−DNA,[52] the GR DBD−DNA,[50] and mutants of the GR DBD−DNA[33] complexes, play an important role in the binding of transcription factor DBDs to DNA. The effects of ionic strength are also neglected here, and it has been demonstrated that this can have a substantial effect on electrostatic interactions in protein−DNA complexes.[8,10] In addition, we neglected the effect of cooperativity of binding of the GR DBD dimer to DNA. In vivo, the GR binds to DNA as a dimer, and Eriksson and Nilsson[33] found that even point mutations in the DBD region determined to be the most important for the specificity of binding (P box) significantly influence the conformations of the residues that form the protein dimer interface in the GRE−(GR $DBD)_2$ complex.

Inclusion of explicit water molecules as well as detailed conformational analysis that would allow better detection of differences in conformational changes upon formation of different complexes would certainly improve our model. Bearing these limitations in mind, we conclude that the specificity of binding of the transcription factor DBDs studied to DNA is largely determined by the energetic cost of DNA desolvation and is tuned by intermolecular electrostatic interactions and conformational changes. The most important interactions are those between mutated bases and the mutated residue (439), the $Zn^{2+}$ ion, and the side chains of charged residues Lys-442, Lys-446, Arg-447, and Arg-470.

**Acknowledgment.** S. Tomic expresses gratitude for a scholarship from the Alexander von Humboldt Foundation. This work was in part supported by the International Buro of the BMBF, Project No. KRO-006-98. We thank Gabriele Cruciani for providing the GOLPE program, Enrica Filipponi for help regarding the use of the GOLPE program and the interpretation of the preliminary results, and Manolo Pastor and Federico Gago for helpful discussions regarding data treatment.

## References

(1) Beato, M. Gene regulation by steroid hormones. *Cell* **1989**, *56*, 335−344.
(2) Evans, R. M. The steroid and thyroid hormone receptor superfamily. *Science* **1988**, *240*, 889−895.

(3) Laudet, V.; Hanni, C.; Coll, J.; Catzefils, F.; Stehelin, D. Evolution of the nuclear receptor gene superfamily. *EMBO J.* **1992**, *11*, 1003−1013.

(4) Mandel-Gutfreund, Y.; Schueler, O.; Margalit, H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **1995**, *253*, 370−382.

(5) Mandel-Gutfreund, Y.; Margalit, H. Quantitative parameters for amino acid−base interaction: implications for prediction of protein-DNA binding sites. *Nucl. Acid Res.* **1998**, *26*, 2306−2312.

(6) Eriksson, M. A. L.; Nilsson, L. Structure, Thermodynamics and Cooperativity of the GR DBD in Complex with Different REs. MD Simulation and Free Energy Pertubation Studies. *J. Mol. Biol.* **1995**, *253*, 453−472.

(7) Komeiji, Y.; Fujita, I.; Honda, N.; Tsutsui, M.; Tamara, T.; Yamato, I. Glycine 85 of the trp-repressor of E.coli is important in forming the hydrophobic tryptophan binding pocket: experimental and computational approaches. *Protein Eng.* **1994**, *7*, 1239−1247.

(8) Zacharias, M.; Luty, B. A.; Davis, M. E.; McCammon, J. A. Poisson−Boltzmann analysis of the lambda repressor-operator interaction. *Biophys. J.* **1992**, *63*, 1280−1285.

(9) Zacharias, M.; Luty, B. A.; Davis, M. E.; McCammon, J. A. Combined conformational search and finite-difference Poisson−Boltzmann approach for flexible docking. Application to an operator mutation in the lambda repressor-operator complex. *J. Mol. Biol.* **1994**, *238*, 455−465.

(10) Misra, V. K.; Hecht, J. L.; Yang, A. S.; Honig, B. Electrostatic contributions to the binding free energy of the lambdacI repressor to DNA. *Biophys. J.* **1998**, *75*, 2262−2273.

(11) Brown, L. M.; Bruccoleri, R. E.; Novotny, J. Empirical free energy calculations of phage 434 repressor- and cro-DNA complexes support the 'indirect readout' hypothesis of specificity. *Pac. Symp. Biocomput.* **1998**, 339−348.

(12) Fogolari, F.; Elcock, A. H.; Esposito, G.; Viglino, P.; Briggs, J. M.; McCammon, J. A. Electrostatic effects in homeodomain-DNA interactions. *J. Mol. Biol.* **1997**, *267*, 368−381.

(13) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA): 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, **1988**, *110*,

(14) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; deSolms, S. J. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **1995**, *38*, 305−17.

(15) Grootenhuis, P. D. J.; van Galen, P. J. M. Correlation of binding affinities with nonbonded interaction energies of thrombin-inhibitor complexes. *Acta Crystallogr.* **1995**, *D51*, 560−566.

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

(17) Wade, R. C.; Ortiz, A. R.; Gago, F. Comparative Binding Energy Analysis. *Perspect. Drug Des. Discov.* **1998**, *9*, 19−34.

(18) Wold, S.; Johansson, E.; Cocchi, M. *PLS−Partial Least-Squares Projections to Latent Structures*; ESCOM: Leiden, 1993.

(19) Zilliacus, J.; Wright, A. P.; Norinder, U.; Gustafsson, J. A.; Carlstedt-Duke, J. Determinants for DNA-binding site recognition by the glucocorticoid receptor. *J. Biol. Chem.* **1992**, *267*, 24941−24947.

(20) Zilliacus, J.; Wright, A. P.; Norinder, U.; Gustafsson, J. A.; Carlstedt-Duke, J. DNA-binding specificity of mutant glucocorticoid receptor DNA-binding domains. *Ann. N. Y. Acad. Sci.* **1993**, *684*, 253−255.

(21) Zilliacus, J.; Wright, A. P.; Carlstedt-Duke, J.; Gustafsson, J. A. Structural determinants of DNA-binding specificity by steroid receptors. *Mol. Endocrinol.* **1995**, *9*, 389−400.

(22) Zilliacus, J.; Wright, A. P.; Carlstedt-Duke, J.; Nilsson, L.; Gustafsson, J. A. Modulation of DNA-binding specificity within the nuclear receptor family by substitutions at a single amino acid position. *Proteins: Struct. Funct. Genet.* **1995**, *21*, 57−67.

(23) Hard, T.; Gustafsson, J. A. Structure and function of the DNA-binding domain og the glucocorticoid receptor and other members of the nuclear receptor supergene family. *Acc. Chem. Res.* **1993**, *26*, 644−650.

(24) Klock, G. S.; U.; Schuetz, G Oestrogen and glucocorticoid response elements are closely related dut distinct. *Nature (London)* **1987**, *329*, 734−736.

(25) Hard, T.; Kellenbach, E.; Boelens, R.; Kaptein, R.; Dahlman, K.; Carlstedt-Duke, J.; Freedman, L. P.; Maler, B. A.; Hyde, E. I.; Gustafsson, J. A.; et al. 1H NMR studies of the glucocorticoid receptor DNA-binding domain: sequential assignments and identification of secondary structure elements. *Biochemistry* **1990**, *29*, 9015−9023.

(26) Tsai, S. Y.; Carlstedt-Duke, J.; Weigel, N. L.; Dahlman, K.; Gustafsson, J. A.; Tsai, M. J.; O'Malley, B. W. Molecular interactions of steroid hormone receptor with its enhancer element: evidence for receptor dimer formation. *Cell* **1988**, *55*, 361−369.

(27) Luisi, B. F.; Xu, W. X.; Otwinowski, Z.; Freedman, L. P.; Yamamoto, K. R.; Sigler, P. B. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **1991**, *352*, 497−505.

(28) Free, S. M.; Wilson, J. A. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(29) Kubinyi, H. *The Free−Wilson method and its relationship to the extrathermodynamic approach*; Pergamon Press: Oxford, 1990; Vol. 4, pp 589−643.

(30) Wold, S.; C.; A.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjoestroem, M. In *Chemometrics-Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; Reidel: Dordrecht, 1984; pp 17−95.

(31) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool For Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.

(32) GOLPE; Multivariate Infometric Analysis S.r.l.: Perugia, Italy, 1997.

(33) Eriksson, M. A. L.; Nilsson, L. Structural and dynamic effects of point mutations in the recognition helix of the glucocorticoid receptor DNA-binding domain. *Protein Eng.* **1998**, *11*, 589−600.

(34) Eriksson, M. A. L.; Nilsson, L. Structural and dynamic differences of the estrogen receptor DNA-binding domain, binding as a dimer and as a monomer to DNA: molecular dynamics simulations studies. *Eur. Biophys. J.* **1999**, *28*, 102−111.

(35) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular enegy, minimization and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(36) Cornell, W. D.; Cieplak, P.; Payly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(37) Besler, B. H.; Merz, K. M. J.; Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(38) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(39) Creighton, T. E. *Proteins: Structures and Molecular Properties*; W. H. Freeman and Co.: New York, 1993.

(40) Fauchere, J. L.; Pliska, V. *Eur. J. Med. Chem.-Chim. Ther.* **1983**, *18*, 369−375.

(41) Wimley, W. C.; Creamer, T. P.; White, S. H. Solvation Energies of Amino Acid Side Chains and Backbone in a Family of Host−Guest Pentapeptides. *Biochemistry* **1996**, *35*, 5109−5124.

(42) Kollman, P. Free Energy Calculations: Application to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.

(43) Shih, P.; Pedersen, L. G.; Gibbs, P. R.; Wolfenden, R. Hydrophobicities of the Nucleic Acid Bases: Distribution Coefficients from Water to Cyclohexane. *J. Mol. Biol.* **1998**, *280*, 421−430.

(44) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.

(45) Hubbard, S. J.; Thornton, J. M. NACCESS; Department of Biochemistry and Molecular Biology, University College London: London, 1993.

(46) INSIGHTII; Molecular Simulations Inc., San Diego, 1997.

(47) Plaxco, K. W.; Goddard, W. A. I. Contributions of the thymine methyl group to the specific recognition of poly and mononucleotides: An analysis of the relative free energies of solvation of thymin and uracil. *Biochemistry* **1994**, *33*, 3050−3054.

(48) Perez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.* **1998**, *41*, 836−852.

(49) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of comparative molecular field analysis models: effects of data scaling and variable selection using a set of human synovial fluid phospholipase A2 inhibitors. *J. Med. Chem.* **1997**, *40*, 1136−1148.

(50) Eriksson, M.; Haerd, T.; Nilsson, L. Molecular Dynamics simulations of the glucocorticoid receptor DNA-binding domain in complex with DNA and free in solution. *Biophys. J.* **1995**, *68*, 402−426.

(51) Williams, D. H.; Maguire, A. J.; Tsuzuki, W.; Westwell, M. S. An analysis of the origins of a cooperative binding energy of dimerization. *Science* **1998**, *280*, 711−714.

(52) Kosztin, D.; Bishop, T. C.; Schulten, K. Binding of the Estrogen Receptor to DNA. The Role of Waters. *Biophys. J.* **1997**, *73*, 557−570.

JM9911175