CROATICA CHEMICA ACTA CCACAA **74** (2) 295–314 (2001)

ISSN-0011-1643
CCA-2733 *Original Scientific Paper*

COMBINE Analysis of Nuclear Receptor-DNA Binding Specificity: Comparison of Two Sets of Data

Sanja Tomić[a,*] *and Rebecca C. Wade*[b]

[a]*Ruđer Bošković Institute, P. O. Box 180, HR-10002 Zagreb, Croatia*

[b]*European Molecular Biology Laboratory, Meyerhofstr. 1, D-69012 Heidelberg, Germany*

Received January 5, 2001; revised February 26, 2001; accepted February 28, 2001

To identify the major determinants of the DNA binding specificity of nuclear transcription factors, the Comparative Binding Energy (COMBINE) analysis has been performed for two datasets. In Tomić *et al*.,[1] COMBINE QSAR models were derived for a set of 320 complexes of DNA and glucocorticoid receptor mutants. Here, we derive COMBINE QSAR models for a set of 32 complexes. This set differs from the larger one in two aspects. The complexes have additional mutation sites in the DNA binding domain and, instead of just activity measurements, both activity and binding affinity measurements are available. Models of better predictive ability were obtained with the smaller, but experimentally better characterized, dataset.

The parameters important for determining binding specificity are nevertheless similar for both datasets: the electrostatic interaction energies between the mutated nucleotides and mutated residue(s) as well as some charged amino acid residues (Arg-447, Arg-470, Arg-477), and the solvation free energies of the mutated base(s). However, the relative importance of these parameters is different in the two datasets.

Key words: Quantitative Structure-Activity Relationship (QSAR), COMBINE analysis, molecular modeling, gene regulation, molecular mechanics.

* Author to whom correspondence should be addressed. (E-mail: tomic@faust.irb.hr)

INTRODUCTION

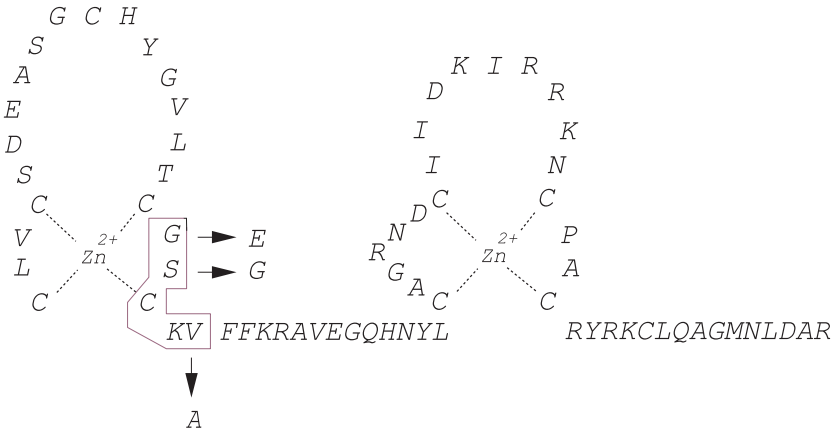
DNA transcription is crucial for gene regulation. In some cases, transcription occurs only upon the binding of a protein, the so-called transcription factor, to DNA. In this work, we study the specificity of the binding of nuclear receptor transcription factors, namely glucocorticoid and estrogen receptors, to DNA. The glucocorticoid receptor (GR) and the estrogen receptor (ER) are steroid hormone receptors and ligand-inducible transcription factors. After the ligand (steroid) binds to the C-terminal ligand binding domain, the conformation of the receptor changes and the DNA binding domain (DBD) becomes exposed and shows a large affinity towards DNA.[2] The GR DBD binds to the glucocorticoid response element (GRE) as a homodimer. The nuclear receptor transcription factor DBDs are highly conserved in sequence and structure and consist of two zinc-fingers (Figure 1a). On binding, part of the recognition α-helix is inserted into the major groove of the DNA. Glucocorticoid and estrogen response elements (ERE) are partially palindromic repeats, consisting of two hexameric half sites with a three-base-pair spacing in-between. They differ from each other in the two central base pairs of each half-site (Figure 1b).

Zilliacus *et al*.[3,4] studied how mutations of a few residues of the GR DBD to residues specific for the ER DBD modulate DNA binding affinity. In Zilliacus *et al*.,[4] they mutated Gly-439 in the GR DBD to the remaining 19 natural amino acid residues and measured the interaction of these 20 DBDs with 16 different response elements by a transactivation assay. We used COMBINE and Free-Wilson analysis[1] to derive QSARs for these 320 complexes. Predictive models were derived by both methods and the COMBINE QSAR model provided insight into the physico-chemical factors determining binding specificity.

Here we describe the COMBINE QSAR models for a different set of 32 complexes studied by Zilliacus *et al*.[3] This dataset consists of 8 different DBDs bound to 4 different response elements and both activity and binding affinity measurements are available. Another difference between these two datasets is the number of variable positions. Instead of only one mutation site in the DBD, there are three mutation sites in the DBD in the smaller dataset. Namely, one to three specificity determining residues in the GR DBD are mutated to the corresponding residues in the ER (Gly-439 to Glu, Ser-440 to Gly and Val-443 to Ala) (Figure 1a).

The aim of this work is to compare the physical parameters determined with the 32-object dataset with those determined earlier[1] with the 320-object dataset, as well as to determine the dependence of prediction quality on the data size, type of experimental measurements and the number of mutation sites in DBD.

a)



b)

GRE	AG**AA**CA
	TC**TT**GT
ERE	AG**GT**CA
	TC**CA**GT
GRE2	AG**GA**CA
	TC**CT**GT
ERE2	AG**AT**CA
	TC**TA**GT

Figure 1. (a) Amino acid sequence of the human GR DBD. Arrows show the amino acid substitutions introduced into the mutant protein. The introduced residues are characteristic of the ER DBD; (b) Sequences of the glucocorticoid and estrogen response elements (RE) and their mutants (GRE2 and ERE2), for which the specificity of binding of native and mutated GR DBDs was measured.[3]

COMPUTATIONAL METHODS

COMBINE Analysis

The COMBINE approach is described in detail elsewhere.[1,5,6] The main aim of this method is to approximate the binding free energy (ΔG), in this case the ΔG of DBD-DNA binding, with the sum of n selected, residue based, terms Δu_i^{sel}:

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

The expression used in this work as an approximation of the DBD-DNA binding affinity is:

$$\Delta G = \sum_{i=1}^{n_{DNA}} \sum_{j=1}^{n_{DBD}} w_{ij} u_{ij}^{vdw} + \sum_{i=1}^{n_{DNA}} \sum_{j=1}^{n_{DBD}} w_{ij} u_{ij}^{ele} + \sum_{j=1}^{N_{DBD}} w_j \Delta\Delta G_j^{hyd} + \sum_{k=1}^{4} w_k \Delta\Delta G_{mb\ k}^{hyd} +$$

$$\Delta E_{DNA} + \Delta E_{DBD} + \sum_{j=1}^{N_{DBD}} (w_j \Delta SA_j^{p} + w_j \Delta SA_j^{np}) + \sum_{i=1}^{N_{DNA}} w_i \Delta SA_i \qquad (2)$$

where u_{ij} is the intermolecular interaction energy between group i in the DNA and group j in the DBD; ΔE_x (X = DNA, DBD) is the change of the potential energy of the corresponding molecule upon binding; $\Delta\Delta G_j$ and $\Delta\Delta G_{mb\ k}$ are the change of relative free energy of solvation of the side chain of residue j and of the mutated base (mb) k, respectively; ΔSA_i is the change of the solvent accessible surface of nucleotide i; ΔSA_j^p and ΔSA_j^{np} are the changes of the polar and nonpolar solvent accessible surface of the amino acid residue j, respectively; n_{DNA} and n_{DBD} are the number of groups in 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.

The change of the solvent accessible surface per residue upon binding (ΔSA) was calculated using the NACCESS program.[7,8] For the calculation of the relative free energy of hydration, empirical values relative to glycine in the pentapeptide AcGG-X_i-GG[9] were used for amino acids while values relative to thymine were used for bases.[10] The calculation details for individual terms were given in our previous work.[1] The weight, w, of each term in Equation 2 was determined by PLS analysis.

Building DNA-DBD Complexes

In the study, systems consisting of the 73 residue mutated GR DBD (Figure 1a) bound to a 6-base pair (bp) mutant of the GR RE were used (Figure 1b). The structures of the mutated complexes were derived from the crystallographically determined structure of the rat GR DBD-DNA complex.[11] The mutants were modeled as follows. One to three of the residues, Gly-439, Ser-440 and Val-443 in the specifically bound protein monomer (DBD1) of the crystal structure, were replaced with the corresponding amino acid side chains from the ER, Glu-439, Gly-440 and Ala-443. The base pairs in the central positions (3 and 4, Figure 1b) of the GR RE were subsequently mutated to those of the ER RE. In this way, 32 different complexes (8 different DBDs bound to 4 REs) were obtained for which experimental data were available.[3]

Modeling was performed with the all-atom AMBER molecular mechanics force field.[12] The parameters for Zn^{2+} were derived from the results of MOPAC 6.0 AM1[13,14] computations and are given in our previous work.[1] The conformation of each DBD, DNA and their complex was minimized with the AMBER 6.0 program[15] using a distance dependent dielectric constant. Four datasets were considered: 32–300, geometry optimized 300 steps with the backbone restrained by a harmonic force, k = 210 kJ mol^{-1} Å^{-2}; 32–300–30, 32–300–300 and 32–300–700 datasets obtained when the 32–300 dataset was further energy minimized without restraints by 30, 300 or 700 steps, respectively. Partition of interaction energies into parts was performed using the ANAL module in the AMBER program. The interaction energy was split into in-

teraction terms between groups: for each amino acid residue, except for the first one, the interaction was split into side-chain and backbone interactions, and each nucleotide interaction into base, phosphate and sugar interactions. The following intermolecular interactions were considered: side chain-base interaction, backbone-base interaction, side chain-phosphate interaction, backbone-phosphate interaction, side chain-sugar interaction, backbone-sugar interaction for each amino acid and nucleotide pair.

To improve the model, the changes of the free energy of solvation of amino acid side chains and of the mutated bases were considered along with the changes of the solvent accessible surface area per residue and nucleotide as terms contributing to binding affinity. Further, the decomposed intramolecular interaction energy of the complex was also included in the analysis, since it was previously found that it has good predictive abilities.[1]

Data Preparation for Chemometric Analysis

The final matrix of X variables (energy terms) contained 32 rows and 10526 columns: 1–5148 for the van der Waals interaction energies between the groups defined as explained above (u_{ij}), 5149–10296 for the electrostatic interaction energies between the groups (u_{ij}); 10297–10367 for $\Delta\Delta G_j$ per amino acid residue; 10368–10371 for $\Delta\Delta G$ per mutated base; 10372–10513 for ΔSA_j^{p} and ΔSA_j^{np}; 10513–10524 for the change of solvent accessible surface per nucleotide; 10525–10526 for the change of potential energy of the DBD and DNA upon binding. In some models, the change of the conformational energy (ΔE_{co}) of the DBD and DNA upon binding was also considered. However, since the latter two terms do not increase the predictive ability of the models, they are not considered any further. When both inter- and intra-molecular interactions were considered, the final matrix of X variables contained 31862 columns.

The 1D $\boldsymbol{Y}$ matrix consists of 32 logarithmic ($-\log_{10}$) values of the measured binding affinities[3] (ba), where ba equals the amount of protein (ng) needed to bind 30% of the total DNA probe; when 50 ng of protein bound less than 30% of the total DNA probe, ba = 50 or 100, for binding more or less than 10% of the probe, respectively.

The data were subjected to PLS[16] and the principal components (PC) analysis with the GOLPE program.[17,18] The following models are distinguished according to the input X-variables: a) I0-only intermolecular interaction energies; b) G-only differences of the free energies of solvation of amino acid residue side chains and mutated bases; c) SA-only changes of the protein and DNA surface area upon binding (used as an approximation of the desolvation free energy); d) ISA-intermolecular interaction energies and the surface area changes upon binding; e) IG-intermolecular interaction energies and the differences in free energies of solvation; f) IGSA intermolecular interaction energies as well as the differences in free energies of solvation and the changes of surface area upon binding. An additional model, which besides intermolecular interactions, also includes intramolecular interaction energies, is designated: II0.

Chemometric Analysis

i) Data pretreatment. The X-variables were pretreated using different procedures: zeroing, minimum standard deviation cutoff and block scaling ($\boldsymbol{X}$-matrix is divided into k blocks of related data, with blocks scaled so that the same total weight is given to each); for a detailed description see Tomić *et al*.[1]

ii) PLS statistical analysis[16] was performed to derive models and determine the variables that are most important for the specificity of binding.

iii) Principal component analysis (PCA) was performed in order to determine the internal structure of *X*-variables, and to classify the complexes by their distribution in the PC-plots.

iv) The quality of the models was evaluated by checking their predictive abilities. For this purpose, internal and external validation was performed. For internal validation: leave one out (LOO) and random groups cross validation using 5 random groups and 20 randomizations were performed. The predictive ability of a model is described with SDEP and Q^2:

$$\mathrm{SDEP} = \sqrt{\sum \frac{(Y - Y')^2}{N}} \tag{3}$$

$$Q^2 = \frac{\sum (Y - Y')^2}{\sum (Y - \langle Y \rangle)^2} \tag{4}$$

where Y is the experimental, Y' is the predicted and $\langle Y \rangle$ is the average experimental value ($-\log_{10}(\mathrm{ba})$), and N is the number of DBD-DNA complexes.

External validation was performed by dividing the data sets into two: a training set and a test set. In addition, the predictive ability of the models was checked in computations with scrambled *Y*-values.

v) In order to extract the most predictive *X* variables, selection was performed by the fractional factorial design FFD strategy as implemented in the GOLPE program.[18]

RESULTS

In this work, we consider a set of 32 DBD-DNA complexes and 28- and 27-object subsets of the complete dataset. The subsets were used in order to check how much the training set reduction influences the derived models and to evaluate their external predictive ability.

In order to determine which of the four differently optimized datasets (see Computational Methods) would be the best for the COMBINE analysis, the model with intermolecular interaction energy terms (I0) was derived for all four. The best models were obtained with the 32-300 and 32-300-30 datasets (Table I). With the LOO validation procedure, a saddle point was noticed in the Q^2 curve for LV = 2 in both cases. After a single FFD variable selection for the two-dimensional model, the model obtained for the 300-30 dataset had better fit and predictive ability (R^2 = 0.68, Q^2 = 0.52) than that for the 300 dataset (R^2 = 0.56, Q^2 = 0.41). Therefore, the structures of the 300-30 dataset were used for all further analyses.

TABLE I

Predictive performance of the I0-COMBINE model for the 32-object dataset optimized to different extents

Optimization steps[a]	Y[b]	LV[c]	SDEP	(Q^2)[d]	SDEC	R^2
300	L	4	0.34	0.54	0.26	0.74
300 + 30	L	5	0.34	0.54	0.25	0.76
300 + 300	L	5	0.35	0.51	0.26	0.74
300 + 700	L	2	0.39	0.38	0.32	0.59

[a] Numbers refer to the number of steps of restrained and unrestrained energy optimization, respectively.

[b] Binding affinity, L = logarithmic (-log10) values of the measured binding affinity.

[c] Optimal number of LVs for the validation procedure used.

[d] Results of validation performed using the random groups procedure (see Computational Methods).

COMBINE Models for the 32-and 27-Object Datasets

The 27-object data set was constructed by ordering the complexes by the decreasing binding affinity and then removing each sixth complex. The 5 extracted objects served as a test set for the models derived from the 27 complexes.

A number of different COMBINE models were derived for 32- and 27-object datasets. Whenever different blocks of *X*-variables were combined, the block scaling procedure was performed. The fitting and internal validation performance parameters of equivalent models are similar for these two datasets (Tables II and III), as is the structure of the models (Figure 2). Before the fractional factorial design (FFD) variable selection, the II0 model (inter and intra molecular interaction terms) has a predictive performance ($SDEP_{LOO}$ = 0.33, LV = 4) comparative to the other models for both the datasets considered. On the other hand, after FFD variable selection, all models including the intermolecular interaction energy terms only (I0, IG, ISA, IGSA) have a better predictive ability than the II0 model ($SDEP_{LOO}$ = 0.32, LV = 2).

I0 Models (Intermolecular Energy Terms Only)

The internal and external predictive ability of the I0 model is good and increases after FFD variable selection, reaching a maximum Q^2 in 2–3 LVs, see Tables II and III. The dominating *X*-variables in the model are the van der Waals and electrostatic interactions between mutated nucleotides (3' and 4'), mostly bases, and the side chains (SC) of mutated amino acid residues (SC439, SC440 and SC443). After FFD, the electrostatic interactions

TABLE II

Predictive performance of COMBINE models derived for the 32-object dataset

Model	Pretreatment[a]	Y[b]	Inter[c]	$(\Delta\Delta G^{hyd})$[d]	ΔSA[e]	LV[f]	SDEP[g]	Q^2	SDEC	R^2
I0	P	L	+			2	0.35	0.51	0.32	0.58
I0	FFD	L	+			2	0.29	0.66	0.24	0.77
I0	–	S	+					<0.0		
SA	P	L			+	4	0.34	0.53	0.28	0.69
SA	FFD	L			+	2	0.34	0.55	0.28	0.68
G	P	L		+		4	0.33	0.57	0.25	0.74
G	FFD	L		+		3	0.27	0.71	0.22	0.81
IG	PB	L	+	+		4	0.33	0.57	0.25	0.74
IG	FFD	L	+	+		3	0.27	0.71	0.22	0.81
ISA	PB	L	+		+	5	0.33	0.57	0.25	0.74
ISA	FFD	L	+		+	2	0.27	0.70	0.22	0.80
IGSA	PB	L	+	+	+	3	0.34	0.52	0.29	0.67
IGSA	FFD	L	+	+	+	1	0.27	0.71	0.23	0.78

[a] Types of X-variable pretreatment; P, zeroing (0.01) and minimum standard deviation cutoff (0.01); B, block scaling; FFD = variable selection by fractional factorial design procedures, 20–50% of dummy variables and combination/variable 2–5 were used. The procedure was repeated until no futher improvement of the predictive performance of a model was detected.

[b] Binding affinity, L = logarithmic (–log10), S = scrambled values of the measured binding affinities.[3]

[c] Intermolecular interaction energies (van der Waals and electrostatic).

[d] Changes of the free energy of solvation (per amino acid residue and mutated base).

[e] Changes of the solvent accessible surface area (per amino acid residue and nucleotide).

[f] Optimal number of LVs. Further increase of the number of LVs does not improve the predictive performance of the model.

[g] Validation is performed by the leave one out procedure.[17,18]

TABLE III

Predictive performance of COMBINE models derived for the reduced 27-object dataset

Model	Pretreatment[a]	Y[b]	Inter[c]	$(\Delta\Delta G^{hyd})$[d]	ΔSA[e]	LV[f]	SDEP[g]	Q^2	SDEC	R^2	SDEP(r)[h]
I0	P	L	+			5	0.34	0.50	0.25	0.74	0.27(0.93)
I0	FFD	L	+			3	0.27	0.68	0.24	0.81	0.24(0.95)
G	PB	L		+		5	0.37	0.43	0.26	0.70	0.30(0.70)
G	FFD	L		+		2	0.35	0.47	0.30	0.61	0.18(0.99)
IG	PB	L	+	+		5	0.35	0.49	0.25	0.73	0.43(0.70)
IG	FFD	L	+	+		3	0.28	0.67	0.21	0.81	0.25(0.93)
IGSA	PB	L	+	+	+	5	0.35	0.49	0.25	0.73	0.28(0.91)[i]
IGSA	FFD	L	+	+	+	4	0.29	0.64	0.22	0.79	0.26(0.94)

[a–g] See corresponding footnotes in Table II.

[h] SDEP value for the external 5-object dataset and correlation (in brackets) between the measured and predicted Y ($-\log_{10}$(ba)) values.

[i] The best external predictivity is obtained using 4 LVs instead of 5.

TABLE IV

Mean[#] predictive performance of COMBINE models derived for the reduced 28-object dataset

Model	Pretreatment[a]	Y[b]	Inter[c]	$(\Delta\Delta G^{hyd})$[d]	LV[f]	SDEP[g]	Q^2	SDEC	R^2	SDEP(r)[h]
I0	P	L	+		4	0.34	0.56	0.25	0.76	0.28 (0.68)
I0	FFD	L	+		4	0.30	0.66	0.22	0.81	0.23 (0.86)
IG	PB	L	+	+	4.3	0.34	0.56	0.25	0.76	0.26 (0.83)
IG	FFD	L	+	+	2	0.29	0.68	0.22	0.81	0.22 (0.89)

[#] Mean values determined considering three different 28-object datasets.

[a-g] See corresponding footnotes in Table II.

[h] Mean SDEP determined for three external 4-object datasets. Values in brackets represent correlation between the experimental and predicted Y for the set of 12 external complexes.

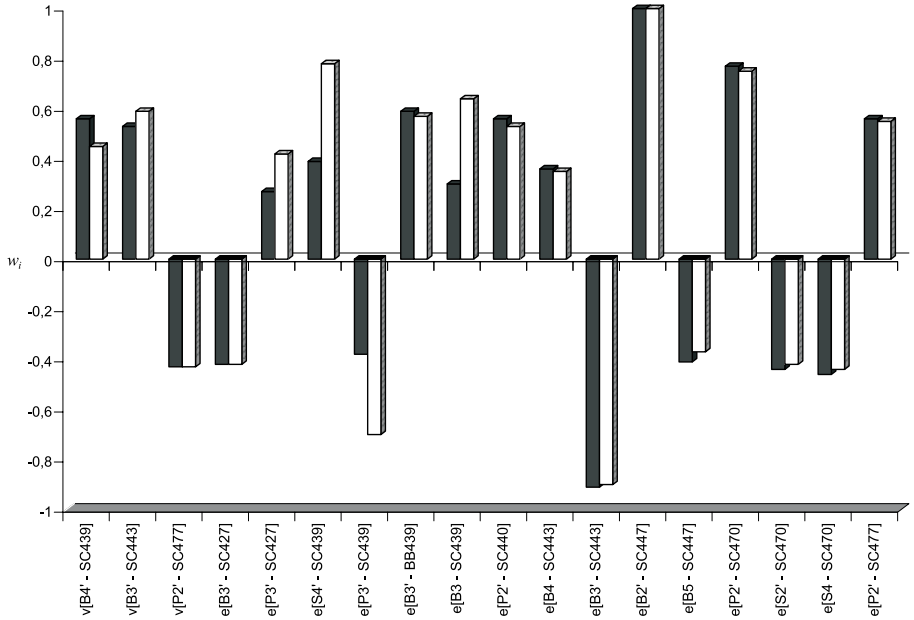


Figure 2. Normalized weighted coefficients for the most important *X*-variables in the I0 model for the 32- (filled) and the 27-object datasets (white) after the FFD variable selection. Normalization is performed with respect to the highest coefficient for each model.

B3'-SC443 and those between nucleotide 2' and the side chains of Arg-447, Arg-470 and Arg-477 appear as very important *X*-variables (Figure 2).

The strong attractive interaction between B2' and Arg-447 as well as those between Arg-470 and Arg-477 and the phosphate group of nucleotide 2' (as seen in Figure 3, they are hydrogen bonded) contribute negatively to the relative binding affinity (Figure 2), *i.e.* higher attraction correlates with lower affinity.

The attractive interaction between base 3' and the side chain of the amino acid residue at position 443 (Figure 2) has a positive influences on binding specificity. This interaction is the strongest in complexes with thymine (T) at position 3' and Val at position 443 whose side chain comes close to nucleotide at 3' upon binding, see Figure 3. Otherwise, it is about zero or slightly repulsive (in complexes with adenine at position 3' and Val at 443). Most of the complexes with Val-443 and T3' have an above average binding affinity and most of the complexes with Val-443 and A3' have a below average binding affinity. An exception is the ESVTT complex, which has a below average binding affinity.[3] The reasons for this are: about 2 kJ mol^{-1} weaker attractive interaction between S2' and the side chain of Arg-470 (due to the influence of

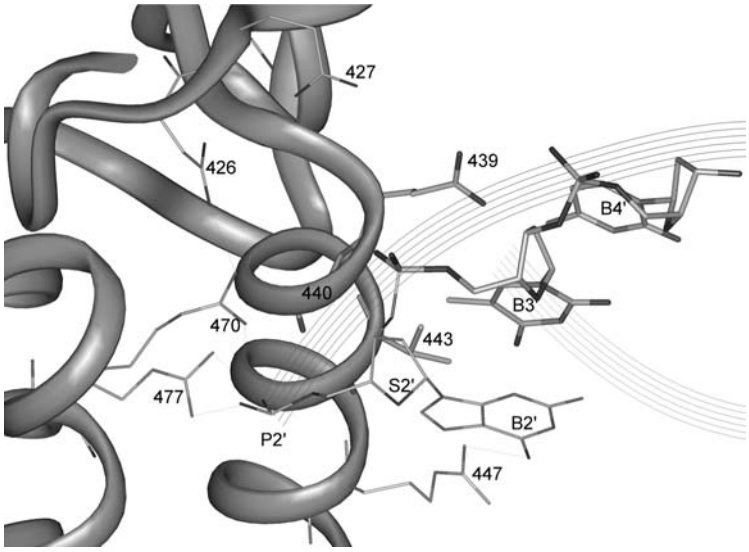


Figure 3. 3D structures of the EGVTC complex. Amino acid residues of the DBD and parts of nucleotides from the RE, which appeared to be the most important structural regulators of protein-DNA binding in the data set considered, are displayed (amino acid residues are labeled by number; parts of nucleotides, phosphate group: P, sugar: S and base: B are designated). Mutated amino acid residues and nucleotides are represented by thicker sticks. Oxygen atoms are colored black, carbons light, and nitrogens dark gray. The protein and DNA backbones are given in ribbon representation (solid and line, respectively).

serine at position 440 on Arg-470) and a slightly stronger repulsion between the side chain of Glu-439 and phosphate group at 3' than in other T-3', Val-443 complexes. On the other hand, the binding affinity of the EGVAC complex is underestimated, probably due to insufficient data in the 32-object data set to correctly predict the SC439-B4' interaction. Whereas in the I0 model derived for the 320-object dataset the relative weight of this interaction is about 10%, in the model derived for the 32-object dataset it is less than 1% of the maximal weight. For the 320 dataset there are 20 different amino acid residues at position 439 while in the 32-object dataset there are only two.

IG, G, ISA, SA, IGSA Models (Combinations of Intermolecular Energy and Free Energy of Solvation and/or Solvent Accessible Surface Area Terms)

The internal and external predictive abilities of these models with block-scaled *X*-variables are similar to that of the I0 model (see Tables II and III). In the IG model, important *X*-variables are almost identical to

those of the I0 model. Besides intermolecular interaction energy terms, the change of free energy of solvation of the amino acid residues 426 and 427 as well as of the base pair 4–3' appear as important *X*-variables. After FFD variable selection, the negative change of the free energy of solvation of the base 3' ($\Delta\Delta G3'$) is strongly correlated with the increasing binding affinity. The model considering the free energy of solvation terms only has the best external predictive ability after FFD variable selection (Table III). The dominating *X* variables in this model are $\Delta\Delta G3'$, $\Delta\Delta G426$, $\Delta\Delta G427$, $\Delta\Delta$G470, *i.e.* *X*-variables which appear as important in the IG model as well. The complexes with the highest binding affinity have thymine (T) at position 3'. Apparently, the desolvation of thymine at position 3' is a driving force in complex formation.

In the ISA and IGSA models, besides the intermolecular interaction energy terms (similar to those in the I0 model), the change of the solvent accessible surface area of residue 439 and the central nucleotides appear as important *X*-variables. After FFD variable selection, the changes of the solvent accessible surface area of residues Asp-426, Glu-427, Arg-470, Lys-471 and Arg-477 and the nucleotide pair 5-2' become more important. However, the decrease of the free energy of hydration of the base at position 3' is an important *X*-variable in the IGSA model, before and after variable selection.

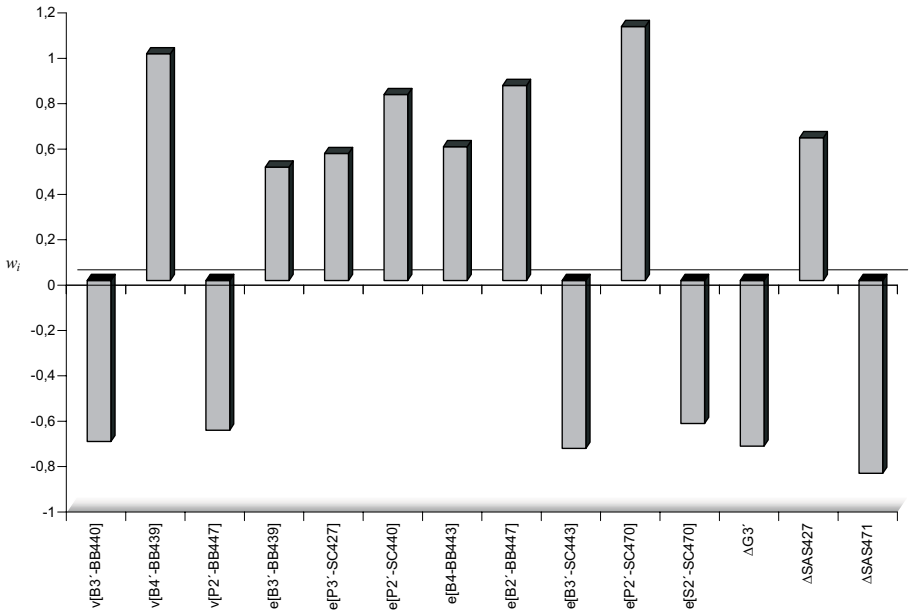


Figure 4. Normalized weighted coefficients for the most important *X*-variables in the IGSA model for the 32-object dataset after the FFD variable selection.

The residues Asp-426, Glu-427, Arg-470, nucleotide 2' and Arg-477 are connected by an H-bond network (Figure 3). Combined investigation of I0, IG and IGSA models as well as 3D structure of the complexes makes the physical explanation of the negative influence of some attractive interactions on the binding affinity clear. Because of the interaction with Ser-440, the side chain of residue 470 is differently oriented in DBDs with serine at position 440 than in those with alanine. Substitution of serine at position 440 causes a twist of the side chain of Arg-470 and exposure of its polar part to water in the free ESA and ESV DBDs (Figure 5).

Desolvation of an arginine is unfavorable and, consequently, the binding of a DBD with Ser-440 to DNA is weakened. But, as the attractive interactions between the phosphate 2' group and SCs of residues 470 and 477 in the

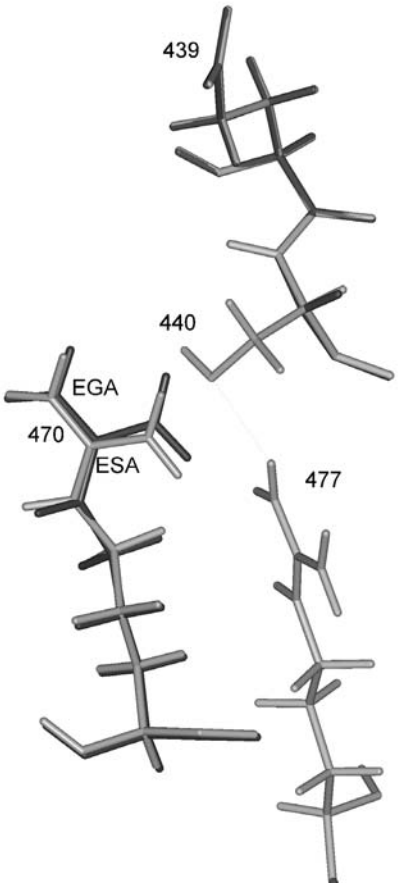


Figure 5. Backbone superimposed 3D structures of the EGA (black) and ESA (gray) complexes. Only residues 439, 440, 470, and 477 are displayed.

complexes with serine at 440 are about 4 kJ mol^{-1} stronger than in those with glycine at 440, they appear with negative weights. Different orientations of the Arg-470 side chain cause differences in the SASA of the amino acid residues with which it is in an H-bond network: Asp-426, Glu-427, nucleotide 2' and Arg-477. Consequently, these SASAs appear as important *X*-variables in the SA, ISA and IGSA models. The model with SASA terms only has lower predictive abilities than the other models (Tables II and III), but still reasonable. In the analysis of intramolecular interactions, we noticed the destabilizing effect of Ser-440 to DBD, resulting from its repulsive interactions with cysteines at positions 438 and 441 of about 12 and 8 kJ mol^{-1}, respectively.

Principal Component Analysis (PC)

Principal components analysis reveals the structure of *X*-variables. In the score plot of the first and second PC of the I0 and IG models, the 32-complexes are divided into four groups with respect to the amino acid residue at position 439 and the nucleotide at position 4' (Figure 6). Distributions of the

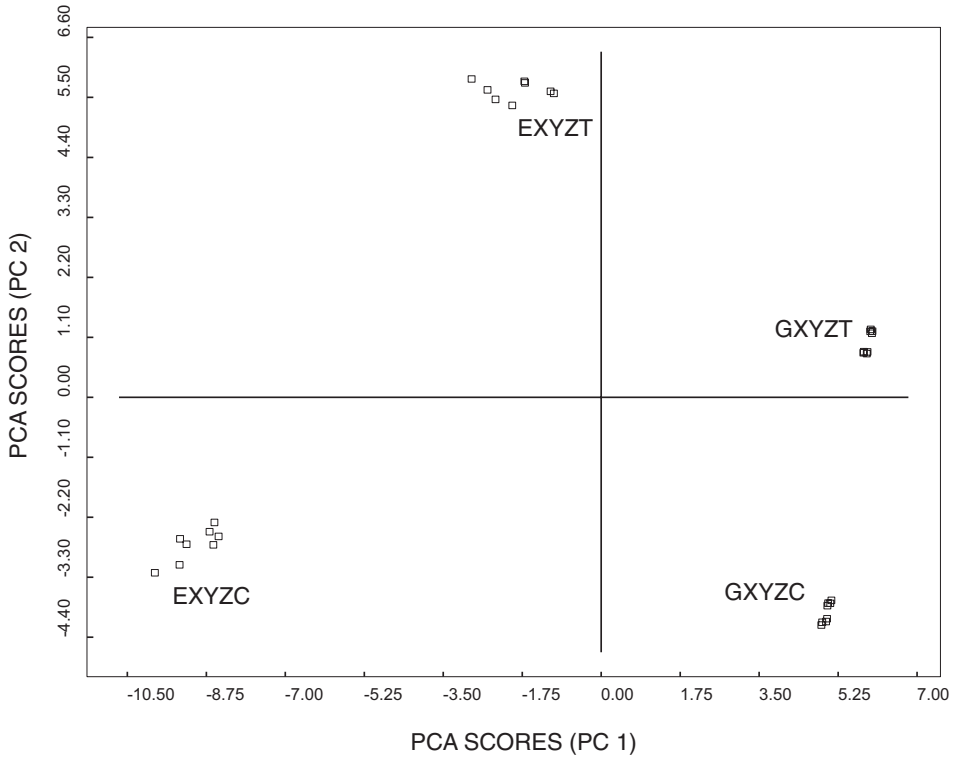


Figure 6. Loadings for the first two PCs of the ISA model.

groups in the score plots of these two models are approximately mirror images on the *y*-axis.

The dominant *X*-variables in both cases are the electrostatic interactions between: (a) the side chain of residue 443 and the base pair 3' 4, and (b) the side chain of residue 439 and the central nucleotides. The electrostatic interaction energy term, SC439-B4', positively describes the first PC in the I0 model, and negatively in the IG model. However, in the latter model, PC1 is mostly described by the $\Delta\Delta G439$ term describing the free energy of hydration of mutated residues (positively). The negative correlation of these two *X*-variables in the IG model and the shift of the electrostatic SC439-B4' interaction term from the right to the left side of the score plot could be the reason why addition of the solvation free energy terms does not improve the predictive ability of the I0 model.

In the score plot of the first and second PCs of the ISA model, the complexes are grouped with regard to the amino acid residues at positions 439 and 443. The dominant *X*-variables are the change of the solvent-accessible surface area of the mutated residues 439 (ΔSA^{p}_{439}) and 443(ΔSA^{np}_{443}).

I0 Model for 28-Object Datasets

Based on the PC analysis, 3 different 28-object datasets were constructed in such a way that one object from each group of the complexes determined by the PC analysis in the I0 and IG models was randomly extracted and put into a 4-object test dataset.

The mean R^2 (SDEC) and Q^2 (SDEP) of the I0 and IG models derived with these datasets as well as external SDEP values are given in Table IV. Prediction of the *Y* variable for the 12 external complexes (3x4) is, after the FFD variable selection, very good (Table IV).

Similarly to models obtained for the whole 32-object dataset, the *Y* value for the EGATC is overestimated (Figure 7). Important *X*-variables in these models are similar to those derived when the 32 and 27 complexes are considered, namely the terms describing the van der Waals and electrostatic interaction energy between mutated amino acid residues and mutated bases, and the electrostatic interaction energy between the central nucleotides and the side chains of residues Arg-447 and Arg-470.

Affinity Prediction for Novel Complexes

On the basis of the analysis of different COMBINE models derived for the 32-object dataset, we tried to predict the sequences of DBDs that would bind to the native GR RE (Figure 1b, central base pair being TT) with high

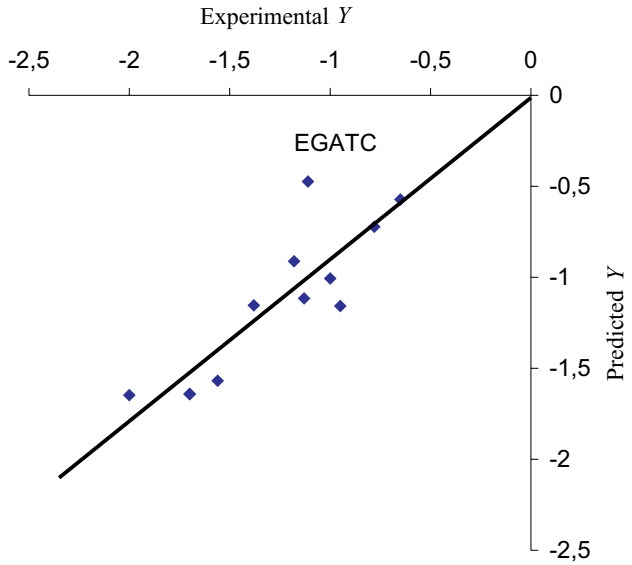


Figure 7. Plot of the predicted against experimental negative $\log_{10}$ binding affinities computed with the I0 model, after the fractional factorial design variable selection was applied, derived with the three different 28-object datasets and applied to the external three 4-object complement datasets.

affinity. Among the 8 new DBDs built, EGI, EGL, EGS, EGT, GGI, GGL, GGS, and GGT, the highest binding affinity was predicted using the I0, IG and G models for GGL, GGS and GGT binding domains. The mean binding affinities (amount of protein in ng needed to bind 30% of total DNA probe) determined for the GGLTT, GGSTT and GGTTT complexes with the I0, IG and G models obtained after the FFD variable selection are 2.3±1.2, 2.5±1.4 and 2.4±1.1, respectively. The binding affinity for the native GR DBD and GR RE is 19 while the highest binding affinity towards GR RE, among the 8 GR DBD mutants, is 4.5 for the GGV and GGA binding domains.[3]

DISCUSSION

In our previous work, we analyzed the binding specificity of transcription factors of the nuclear receptor family to DNA for a set of 320 complexes consisting of all possible combinations of 20 different DBDs with one variable position binding to 16 different response elements with two variable positions.[1] Free-Wilson-like and COMBINE QSAR analyses were performed on this set of 320 complexes using measured functional data from a transactiva-

tion assay. In this work, we used a different dataset of 32-complexes (consisting of 8 different DBDs with three variable positions binding to 4 different response elements with two variable positions), for which binding affinities were available. Using binding affinities, it was possible to derive models with a better predictive ability, despite having fewer complexes in the training set.

The best Q^2 values for the 32 complexes dataset are about 0.7 whereas those for the 320 complexes are about 0.5. The prediction abilities of the smaller dataset are most likely due to the improved accuracy of the binding affinity measurements compared to the functional assay.

The I0, ISA and IGSA models derived with these two different datasets have many common important *X*-variables: the electrostatic energy terms describing the intermolecular interaction between the mutated nucleotides and the mutated residue 439. The terms including the nucleotide pair 5–2', namely the electrostatic interactions of these nucleotides with the side chain of the charged amino acid residues at positions 447, 470 and 477 and the change of their solvent accessible surface area. Thymine at position 3' enhances binding in both datasets because of the favorable desolvation of its methyl group while in the complexes with Val at position 443 because of the attractive CH···O interactions between B3' and the Val-443 side chain (Figure 3). However, all complexes from the 320-object dataset have alanine at position 443. Not all attractive interactions improve binding specificity, *e.g.* the electrostatic interactions B2'-SC447, P2'-SC470 and P2'-SC477 have a negative influence on the relative affinity. Differences in these interactions are mostly due to different bases at position 3' and amino acid residues at position 440. Ser-440 enhances these attractive interactions and in this way indirectly decreases the protein-DNA binding affinity. Mutation of this residue to Gly improves binding affinity in most cases.

The best models with the 320-object dataset were the II0 and IGSA models. However, addition of the solvent accessible surface area and the free energy of solvation to *X*-variables does not enhance the predictive ability of the I0 model with the 32-object dataset although the model, including the free energy of solvation terms only, has a very good predictive ability (Tables II and III), especially external. The negative correlation between the free energy of solvation of the amino acid residue at position 439 and the electrostatic interaction between SC439 and B4', revealed by the principal components analysis, might be a possible explanation for this.

Despite the similarity of the I0 models derived for these two datasets, an attempt to predict activities of the 320 complexes used in our previous work by the model (I0) derived with the 32-object dataset failed. The two main reasons for this failure of prediction are: (a) the different nature of *Y*-vari-

able (binding affinity used in this work and functional activities in the previous work); (b) the different number of variable positions in DBD. According to the results obtained for the 32-object data set, it seems that all three positions in DBD, 439, 440, and 443, are important for defining specificity towards different response elements.

Based on the COMBINE models derived for 32 complexes, a few new DBDs, expected to have a very high affinity towards native GR RE, have been built.

CONCLUSION

The specificity of binding of transcription factors of the nuclear receptor family to DNA is governed by a number of different processes: desolvation of both proteins and DNA, their stability and intermolecular interactions, mostly between charged amino acid residues and phosphate groups of DNA and a few amino acid residues and nucleotides that come in close contact. The specificity of binding can be predicted only within a similar class of complexes (the same variable positions). However, such predictions are reliable (Tables III and IV) and give hints of how to improve binding affinity.

The COMBINE models with the best predictive abilities were derived for the sets of DBD-DNA complexes optimized with a restrained backbone, followed by a small number of unrestrained optimization steps (30 and 70).

In the 320-object dataset studied earlier, the II0 and IGSA models had the best predictive ability. The model with the highest predictive ability (internal and external) for the dataset considered in this work is the I0 model. Its structure is similar to that derived earlier for the 320-object dataset but it has a much better predictive performance. Unlike the 320-object dataset, ΔSASA and $\Delta\Delta G^{hyd}$ do not significantly improve the predictive ability of the I0 model. Although the structures of the models derived earlier and in this work differ, there are a number of important *X*-variables common to both datasets, such as the electrostatic interactions between the bases at positions 3' and 4' and SC439, and base 2', and mutated bases, and the side chains of charged amino acid residues at positions 427, 447, 470 and 477.

Decrease of the free energy of hydration of the B3' base was in both cases found to be the driving force for the formation of the initial complex.

Acknowledgments. – S. Tomić gratefully acknowledges the continuing financial support from the Alexander von Humboldt Foundation. This work was supported in part by the International Bureau of the BMBF, Project No. KRO-006-98. We thank Gabriele Cruciani for providing the GOLPE program.

REFERENCES

1. S. Tomić, L. Nilsson, and R. C. Wade, *J. Med. Chem*. **43** (2000) 1780–1792.
2. P. Arányi, in: M. Simonyi (Ed.), *Problems and Wonders of Chiral Molecules,* Akadémiai Kiadó, Budapest, 1990, pp. 33–44.
3. J. Zilliacus, A. P. Wright, U. Norinder, J. A. Gustafsson, and J. Carlstedt-Duke, *J. Biol. Chem.* **267** (1992) 24941–24947.
4. J. Zilliacus, A. P. Wright, J. Carlstedt-Duke, L. Nilsson, and J. A. Gustafsson, *Proteins: Struct. Func. Genetics* **21** (1995) 57–67.
5. A. R. Ortiz, M. T. Pisabarro, F. Gago, and R. C. Wade, *J. Med. Chem.* **38** (1995) 2681–91.
6. R. C. Wade, A. R. Ortiz, and F. Gago, *Perspect. Drug Des. Discov.* **9** (1998) 19–34.
7. B. Lee and F. M. Richards, *J. Mol. Biol.* **55** (1971) 379–400.
8. S. J. Hubbard and J. M. Thornton, NACCESS, Department of Biochemistry and Molecular Biology, University College London, London, 1993.
9. W. C. Wimley, T. P. Creamer, and S. H. White, *Biochemistry* **35** (1996) 5109–5124.
10. P. Shih, L. G. Pedersen, P. R. Gibbs, and R. Wolfenden, *J. Mol. Biol.* **280** (1998) 421–430.
11. B. F. Luisi, W. X. Xu, Z. Otwinowski, L. P. Freedman, K. R. Yamamoto, and P. B. Sigler, *Nature* **352** (1991) 497–505.
12. W. D. Cornell, P. Cieplak, C. I. Payly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, *J. Am. Chem. Soc.* **117** (1995) 5179–5197.
13. B. H. Besler, K. M. Merz, and P. A. Kollman, *J. Comput. Chem.* **11** (1990) 431–439.
14. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.* **107** (1985) 3902–3909.
15. AMBER 6 program, University of California, San Francisco, 1999.
16. S. C. A. Wold, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjoestroem, in: B. R. Kowalski (Ed.), *Chemometrics-Mathematics and Statistics in Chemistry*, Reidel, Dordrecht, 1984, pp. 17–95.
17. M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, and S. Clementi, *Quant. Struct.-Act. Relat.* **12** (1993) 9–20.
18. GOLPE, University of Perugia, Italy, 1997.

SAŽETAK

Analiza specifičnosti vezanja nuklearnih receptora za DNK COMBINE metodom: usporedba dvaju skupova

Sanja Tomić i Rebecca C. Wade

Da bi se utvrdile osnovne odrednice specifičnosti vezanja nuklearnih transkriptorskih faktora i DNK provedena je komparativna analiza veznih energija (COMBINE).

U prethodnom radu (vidi S. Tomić *et al*.[1]) COMBINE QSAR modeli izvedeni su za skup od 320 kompleksa DNK s mutantima glukokortikoidnih receptora. U ovom radu modeli su izvedeni za skup od 32 kompleksa koji se od većega razlikuje po tomu

što kompleksi imaju dodatno mjesto mutacije u domeni koja se veže za DNK, a pored funkcionalne aktivnosti izmjereni su i afiniteti vezanja. Uporabom manjega, ali eksperimentalno bolje određenog skupa, dobiveni su modeli s boljom pretkaznom moći. Specifičnost vezanja u oba je slučaja bila određena sličnim parametrima, ali različitih relativnih važnosti: energijama elektrostatske interakcije mutiranih nukleotida s mutiranim i nekim nabijenim aminokiselinskim ostatcima (Arg-447, Arg-470, Arg-477), te Gibbsovim slobodnim energijama otapanja mutiranih baza.