June, 1999

New Developments in a Hazard Identification Algorithm for Hormone Receptor Ligands

Steven P. Bradbury
Ovanes Mekenyan
Nina Nikolova, *Bulgarian Academy of Sciences*
Stoyan Karabunarliev
Gerald T. Ankley, et al.

New developments in a hazard identification algorithm for hormone receptor ligands

Ovanes Mekenyan[1], Nina Nikolova[2], Stoyan Karabunarliev[1], Steven P. Bradbury[3]*, Gerald T. Ankley[3] and Bjorn Hansen[4]

[1]Bourgas University "As. Zlatarov", Laboratory of Mathematical Chemistry, 8010 Bourgas, Bulgaria, Bulgarian Academy of Sciences
[2]Central Laboratory for Parallel and Distributed Processing, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
[3]U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Mid-Continent Ecology Division, 6201 Congdon Boulevard, Duluth, MN 55804 USA
[4]Joint Research Centre, Environment Institute, European Chemicals Bureau, I-21020 Ispra (VA), Italy

Abstract

Recently we described the Common REactivity PAttern (COREPA) technique to screen data sets of diverse structures for their ability to serve as ligands for steroid hormone receptors [1]. The approach identifies and quantifies similar global and local stereoelectronic characteristics associated with active ligands through a comparison of energetically-reasonable conformer distributions for selected descriptors. For each stereoelectronic descriptor selected, discrete conformer distributions from a training set of ligands are evaluated and parameter ranges common for conformers from all the chemicals in the training set are identified. The use of discrete partitions of parameter ranges to define common reactivity patterns can, however, influence the outcome of the algorithm. To address this limitation, the original method has been extended by approximating continuous conformer distributions as probability distributions. The COREPA-Continuous (COREPA-C) algorithm assesses the common reactivity pattern of biologically-similar molecules in terms of a product of probability distributions, rather than a collection of common population ranges determined by examination of discrete partitions of a distribution. To illustrate the algorithm, common reactivity patterns based on interatomic distance and charge on heteroatoms were developed and evaluated using a set of 28 androgen receptor ligands. Notable attributes of the COREPA-C algorithm include flexibility in establishing stereoelectronic descriptor criteria for identifying active and nonactive compounds and the ability to quantify three-dimensional chemical similarity without the need to predetermine a toxicophore or align compounds(s) to a lead ligand.

* To receive all correspondence

Key words: structure activity relationships, active analogues, conformational flexibility, androgen receptor binding affinity

Abbreviations:
COREPA-C Parameters
AP, Active pattern, i.e., the reactivity pattern based on the learning set of active chemicals; AT, Active threshold, used to define the "learning" set of active chemicals; Cut-off (AP/NAP), The portion of the nonactive pattern exceeded by the maximum of the active pattern (3-D similarity measure between reactivity patterns of active and nonactive chemicals); D, Euclidean distance; a measure of dissimilarity to compare chemical conformer distributions to training set distributions or to active and nonactive patterns and to compare active and nonactive training set distributions; Γ (gamma), Corresponds to the half-width of a gamma function; $\max P_k^A(x)$, Value of parameter x with the maximum probability of occurrence based on the distribution of the training set of active chemicals; $\max P_k^{NA}(x)$, Value of parameter x with the maximum probability of occurrence based on the distribution of the training set of nonactive chemicals; NAP, Nonactive pattern, i.e., the reactivity pattern based on the learning set of nonactive chemicals; NAT, Nonactive threshold, used to define the "learning" set of nonactive chemicals; $P_i(x)$, Gamma function (probabilistic) distribution of I-th conformer across the axis of molecular descriptor x (single or multiple descriptor values are associated with the conformer when x is a global or local molecular parameter, respectively); $P_k(x)$, Gamma function distribution for all conformers of chemical k across the axis of the molecular descriptor x; $P_k^A(x)$, Probabilistic distribution for the training set of active chemicals across the axis of the molecular descriptor x; $P_k^{NA}(x)$, Overall probabilistic distribution for the training set of nonactive chemicals across the axis of the molecular descriptor x; S(AP/NAP), 3-D similarity between reactivity patterns of active and nonactive chemicals with respect to the molecular descriptor x (overlap of distributions of chemicals within each training set); S(k/AP), 3-D similarity between a chemical k and the active pattern with respect to the molecular descriptor x (overlap between conformer distribution of the chemical and the distribution of the chemicals from the active training set); S(k/NAP), 3-D similarity between a chemical k and the nonactive pattern with respect to the molecular descriptor x (overlap between conformer distribution of the chemical and the distribution of the chemicals from the nonactive training set); $S_{yz}(x)$,

  0931-8771/99/0510-0139 $17.50+.50/0

3-D similarity between two chemicals y and z with respect to the molecular descriptor x (i.e., the overlap between conformer distributions of two chemicals across x).

QSAR Descriptors
AcceptorDeloc-all, Acceptor delocalizabilities of all atoms; Charge-all, Charges of all atoms; Distance-all, Interatomic distance between all atoms; DonorDeloc-all, Donor delocalizabilities of all atoms; d(O_O), Interatomic distances between oxygen atoms; d(R_R), Interatomic distances between all heteroatoms; Egap, Electronic gap ($E_{HOMO} - E_{LUMO}$); E_{HOMO}, Energy of Highest Occupied Molecular Orbital; E_{LUMO}, Energy of Lowest Unoccupied Molecular Orbital; Electronegativity, $0.5(E_{LUMO} + E_{HOMO})$; Geom Wiener, Sum of geometric distances; Max Distance, The greatest interatomic distance; pK_i, Androgen receptor (AR) binding dissociation constant; Planarity, The normalized sum of torsion angles in a molecule; Pol(O), Polarizability of oxygen atoms; Pol(R), Polarizability of all heteroatoms; QH(O), Frontier charges of oxygen atoms on HOMO; QH(R), Frontier charges of all heteroatoms on HOMO; QL(O), Frontier charges of oxygen atoms on LUMO; QL(R), Frontier charges of all heteroatoms on LUMO; Q(O), Charges of oxygen atoms; Q(R), Charges of all heteroatoms; SE(O), Donor delocalizabilities of oxygen atoms; SE(R), Donor delocalizabilities of all heteroatoms; SN(O), Acceptor delocalizabilities of oxygen atoms; SN(R), Acceptor Delocalizabilities of all heteroatoms; Vol.P., Volume Polarizability

1 Introduction

Recently, we reported the development of a technique to systematically assess conformational flexibility in an active analogue search algorithm [1,2]. The COmmon REactivity PAttern (COREPA) approach circumvents the problem of conformer alignment and selection, as well as initial assumptions concerning toxicophore definition [1]. In the algorithm, structural flexibility is incorporated by assessing distributions of all energetically-reasonable conformers of the chemicals under investigation. The COREPA approach was developed, in part, because use of lowest energy conformers for flexible structures to assess similarity in toxicophore search and receptor-mapping algorithms seemed inappropriate. In complex systems such as biological tissues and fluids, it is quite possible that the minimal energy conformer does not interact with the receptor [3] and solvation and/or binding interactions could compensate for energy differences among the conformers of a chemical [4–11].

As originally reported, the COREPA approach incorporates three basic steps [1]. First, training sets for active and nonactive compounds for the chemical series under investigation are selected. This initial step establishes the extent of biological similarity among the chemicals for which stereoelectronic similarity will be defined. In the second step, a restricted set of parameters hypothesized to be associated with biologically similar compounds are evaluated based on a normalized sum of dynamic similarity indices [2]. The similarity method is applied with geometrically and energetically reasonable conformers generated using the method of Ivanov *et al*. [12]. The stereoelectronic parameters that provide the maximal similarity among the chemicals within the training sets, and maximal disimilarity between sets, are assumed to be most closely related to the biological endpoint of concern. In the third step of the algorithm, conformer distributions of the chemicals from the training set are superimposed and the parameter ranges populated by conformers from all the chemicals identified. The collection of common stereoelectronic parameter ranges defines the common reactivity pattern. We initially evaluated this algorithm by defining the stereoelectronic requirements associated with the binding affinity of a diverse set of 28 ligands to the androgen receptor (AR) [1].

While the development and initial evaluation of COREPA was successful, three potential limitations were noted, due solely to the method used to characterize the reactivity patterns. First, we observed that the reactivity patterns could be affected by the extent to which the parameter distributions were partitioned into discrete intervals. Second, similarity assessments between molecules could be strongly biased when there were large differences in conformational flexibility among the chemicals. Finally, a quantitative analysis of similarity between a conformer distribution for an individual chemical and a reactivity pattern derived from a training set was not possible with discrete partitioning of parameter distributions.

To address these limitations, the objective of the present study was to further develop the COREPA approach by establishing a technique to describe conformer distributions with a gamma function. This continuous, rather than discrete, approach to defining conformer distributions facilitates the definition of reactivity patterns as a product of probabilistic distributions rather than a collection of population ranges. To illustrate the capability of the continuous version of COREPA (COREPA-C) we re-evaluated the set of 28 steroidal and non-steroidal AR ligands used previously to assess the discrete COREPA algorithm version.

2 Methods

The basic assumptions and methodology of the COREPA approach are described elsewhere [1,5] and readers are encouraged to consult these references for a detailed presentation of principles and techniques. Briefly, the elucidation of chemical similarity with the COREPA approach is based on the assumption that chemicals that elicit similar biological behavior through a common mechanism of action should possess a commonality in stereoelectronic descriptors. Determination of this common

reactivity pattern within a set of toxicologically-similar chemicals includes examination of the conformational flexibility of the compounds to evaluate molecular similarity in the context of the associated variability in specific stereoelectronic parameters.

2.1 The COREPA-C Approach

2.1.1 Background

For each stereoelectronic parameter (denoted as $\underline{x}$), the distributions of energetically reasonable conformers of the compounds from a training set are approximated by continuous probabilistic functions. First, a probabilistic distribution, $P_i(x)$, across $\underline{x}$ is created for each conformer $\underline{i}$ (and, if the descriptor is a local index, for each atom or bond of the conformer):

$$P_i(x) = (1/n)\sum_{j=1}^{n} P'_j(x_j) \quad (1)$$

where x_j are the discrete values of the parameter $\underline{x}$. Each discrete value x_j of $\underline{x}$ is approximated by a gamma function, $P_j'(x_j)$, defined as follows:

$$P'_j(x_j) = (\Gamma/2)^2/[(\Gamma/2)^2 + (x_j - x)^2] \quad (2)$$

where Γ (gamma) corresponds to the width of the gamma function. This parameter contributes to the shape of the distribution curve and defines the precision of the continuous approximation. As described in Equation 1, $P_i(x)$ is normalized over the number of x_j values. It should be noted that other probabilistic functions could be used. For example, COREPA-C analyses based on a gaussian function, rather than a gamma function, provided similar results (data not shown).

The overall probabilistic distribution, $P_k(x)$, for a compound $\underline{k}$, is then approximated by summing the distributions of its conformers, with the resulting summation normalized by the number of conformers, $\underline{m}$, of the compound:

$$P_k(x) = (1/m)\sum_{i=1}^{m} P_i(x) \quad (3)$$

The overall distributions for the training sets of active (A) and nonactive (NA) chemicals are obtained as a product of the corresponding compound distributions:

$$P_k^A(x) = (1/C_1)\prod_{k=1}^{A} P_k(x) \quad (4)$$

$$P_k^{NA}(x) = (1/C_2)\prod_{k=1}^{NA} P_k(x) \quad (5)$$

where C_1 and C_2 are constants used to normalize the distributions to a total probability of 1.0.

An examination of Eqs. 4 and 5 indicates distribution densities will be high for training sets of chemicals when their conformer distributions are "in-phase." In addition, the training set distribution will be strongly influenced by those compounds whose conformers are most in-phase. Conformer distributions that are dissimilar from the predominant distributions tend to be canceled in the generation of the overall distribution density. Consequently, the overlap of an overall pattern distribution for active chemicals, with the distribution of an individual active chemical that was less in-phase than other members of the training set could result in a nonactive classification. At the same time, it is possible that the distribution of a nonactive chemical could overlap with the active pattern, but still not be in-phase with the distribution of conformers from the active training set. We are currently exploring an alternative version of Eq. 5, where the overall distribution of nonactive or active chemicals is obtained by summing the corresponding compound distributions; e.g., for nonactive chemicals:

$$P_k^{NA}(x) = (1/C_2)\sum_{k=1}^{NA} P_k(x) \quad (6)$$

Essentially, these summed distributions are wider than those derived in Eqs. 4 and 5. As a result, distributions based on sums, rather than ranges obtained by products, could lead to an increase number of false negative identifications of active compounds. There is not an overt mechanistic rationale for using either type of distribution; however, the user can assess the likelihood for false positive and false negative identification of active compounds by comparing the two types of distributions.

Within the COREPA-C approach, the similarity, S_{yz}, between chemicals $\underline{y}$ and $\underline{z}$, with respect to molecular parameter $\underline{x}$, is defined as the percentage of the total area of the distribution curve for chemical $\underline{y}$ that overlaps with the distribution for chemical $\underline{z}$:

$$S_{yz}[\%] = P_y(x) \cap P_z(x) \quad (7)$$

Because the probabilistic distributions are normalized to unity, by definition, S_{yz} is equal to S_{zy}. Thus, within COREPA-C, a single estimate is used to assess similarity between two molecules in terms of the specific descriptors hypothesized to be associated with the biological endpoint under investigation. In the previous version of the algorithm [1], four indices were employed to assess pair-wise similarity among the descriptors of interest.

The joint analysis of reactivity patterns for sets of active and nonactive chemicals allows one to assess the degree of

similarity between active and nonactive patterns (APs and NAPs, respectively):

$$S(AP/NAP)[\%] = P^{A}(x) \cap P^{NA}(x) \quad (8)$$

Moreover, the portion of the nonactive pattern exceeded by the maximum of the active pattern (i.e., Cut-off(AP/NAP) [%]) can be used as a measure of similarity between the active and nonactive patterns. The Cut-Off(AP/NAP) corresponds to the probability that conformers of nonactive chemicals would be identified as similar to conformers of the active chemicals.

Indices for measuring similarity between a chemical's distribution and active or nonactive patterns are described in Eqs. 9 and 10:

$$S(k/AP)[\%] = P_{k}(x) \cap P^{A}(x) \quad (9)$$

$$S(k/NAP)[\%] = P_{k}(x) \cap P^{NA}(x) \quad (10)$$

S(k/AP) and S(k/NAP) can be used to assess the ability of the patterns to discriminate among chemicals. For example, higher values of S(k/AP) for active chemicals and lower S(k/AP) values for nonactive chemicals suggest the active pattern discriminates the compounds in active and inactive training sets.

To further evaluate the dissimilarity between overall patterns of active and nonactive chemicals, as well as between overall patterns and chemical-specific distributions, a Euclidean distance metric (D) was employed. This distance metric was based on the squared differences between distribution densities ($P_i(x)$) over the entire range of the parameter $\underline{x}$:

$$D(AP/NAP) = \sqrt{\int [P^{A}(x) - P^{NA}(x)]^2 dx} \quad (11)$$

$$D(k/AP) = \sqrt{\int [P^{k}(x) - P^{A}(x)]^2 dx} \quad (12)$$

$$D(k/NAP) = \sqrt{\int [P^{k}(x) - P^{NA}(x)]^2 dx} \quad (13)$$

A D value = 0 would indicate that two distributions entirely coincide. Thus, smaller D values are associated with similar distributions. The Euclidean distance metric can also be used to compare distributions derived from different chemical training sets or for training sets derived from different weighting schemes (e.g., different $\Delta\Delta H_f^o$ thresholds for conformer selection).

The Euclidean distance metric can also be used to ascertain the extent to which an overall conformer distribution of active or nonactive chemicals is influenced by a specific compound(s). The pattern "stability" in this respect can be assessed by a "leave-one-out" procedure. The Euclidean distance metric is used to iteractively assess differences between patterns derived for n vs. n-1 chemicals in the training subsets. Variation of similarity indices, cut-offs between active and nonactive patterns, and parameter ranges can also be quantified. Smaller Euclidean distances, variations in similarity indices and corresponding parameter ranges, are associated with more stable patterns.

The COREPA approach requires distributions of all energetically-reasonable conformers to be analyzed. As described in *AR Ligands: Binding Affinity and Electronic Structure*, a 20 kcal/mol threshold for $\Delta\Delta H_f^o$ is assumed to result in an energetically-reasonable set of conformations in the context of ligand-receptor interactions. Of course, reactivity patterns can be derived for any threshold of $\Delta\Delta H_f^o$ thought to be appropriate for the process being modeled. In addition, conformers of each chemical can be considered as a statistical ensemble, based on the Boltzman's statistics. Thus, a conformer weighting scheme could be based on the calculated heats of formation for each ith conformer, ΔH_f^{oi}, which is compared exponentially to the conformer associated with the absolute energy minimum, ΔH_f^{ol} (at $T = 298^{\circ}K$). The statistical weight of the ith conformer is calculated as:

$$p_i = \exp -[\Delta H_f^{oi} - \Delta H_f^{ol})/RT]/\Sigma \exp -[\Delta H_f^{oi} - \Delta H_f^{ol})]/RT \quad (14)$$

In the present work, however, conformer distributions were derived with equally weighted conformers, because of uncertainty in applying a gas-phase energetic assessment to the AR ligand binding domain.

The reactivity patterns obtained by COREPA-C are described by a collection of parameter ranges, whose widths depend upon the values of Γ and/or confidence limits selected for the pattern probability maxima. The continuous approximations of a conformer distribution is less precise as Γ increases, which lowers the intensity of $P_k(x)$ and ultimately leads to larger parameter ranges. Alternatively, with a decrease in Γ, the approximation of a continuous distribution becomes more precise, which increases the intensity of $P_k(x)$ and results in a smaller parameter range; i.e., the portion of the distribution area around the maximum decreases for a specified confidence limit. These influences on $P_k(x)$ are illustrated in Figure 1, which demonstrate increased resolution of distribution patterns with a decrease in Γ (also refer to Figure 5a). Continued lowering of Γ eventually converts COREPA-C into the discrete version of the approach [1]. Increasing Γ will ultimately widen

distributions to the point that discrimination of reactivity patterns is no longer possible because of the extensive overlap between active and nonactive patterns. Thus, the selection of Γ and the confidence interval around a maximum of a pattern probability provide flexibility in describing reactivity patterns. For example, increasing Γ values for active distributions (i.e., precision of the continuous approximation decreased), results in larger parameter intervals and less restrictive parameter screens for active compounds. This leads to an increasing probability of an unknown compound being defined as similar to the active pattern, compared to a pattern derived with a lower Γ. Consequently, increasing Γ can increase the rate of false positive identifications, but lower the rate of false negative identifications. These characteristics may be desired in a screening-level risk assessment of a large database of unknown chemicals, where a low rate of false negatives is required, and an elevated rate of false positives is acceptable.

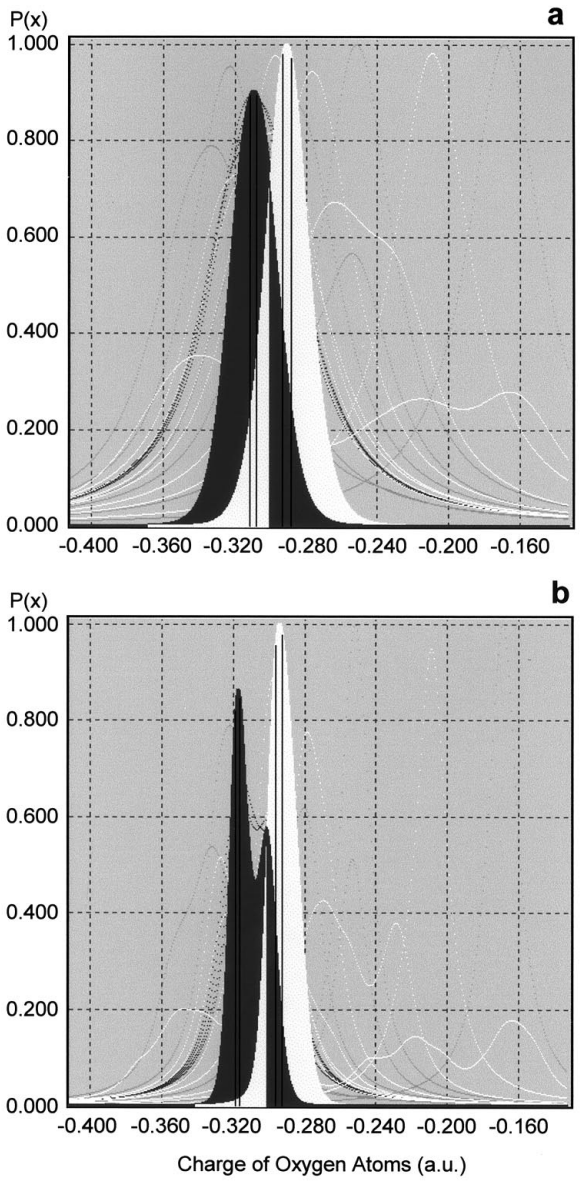

Figure 1. Reactivity patterns of active ($pK_i \geqslant 1.0$) and nonactive ($pK_i \leqslant -2.0$) androgen receptor ligands based on charge on oxygen atoms using $\Gamma = 0.05$ (a) or $\Gamma = 0.02$ (b). The common distribution curves for active chemicals are red, while the distribution for nonactive chemicals is white. Distributions from ligands with pK_i values between 0.7 and −2.0 are colored in green.

In subsequent analyses, Γ values were generally in the range of 0.005 to 0.01 for distributions based on atomic electronic indices, which is similar to semi-empirical quantum-chemical calculation accuracy. For global parameters, Γ values of 0.125 to 0.1 of the parameter variation were generally selected. The selection of Γ values was based on an analysis that related Γ to S(AP/NAP), Cut-off(AP/NAP) and the width of distributions, as represented by confidence limits around pattern probability maxima values, for this data set. It is important to note that relationship of Γ values to S(AP/NAP), Cut-off(AP/NAP) and the width of distributions is highly dependent on the nature of the chemicals in a training set and the descriptors employed.

2.1.2 The Algorithm

The steps in the COREPA-C algorithm are similar to those in Mekenyan *et al.* [1]; only new aspects of the analysis are highlighted.

Step 1. Definition of the training set of chemicals. Two subsets of chemicals from the data set under investigation are selected as the training sets. The first subset consists of chemicals having activity above a user-defined threshold (the active threshold, AT). The second subset includes chemicals having activity not exceeding a user-imposed threshold (the nonactive threshold, NAT). This initial step defines biological similarity among the chemicals in the respective training sets. The stereoelectronic similarity among the chemicals within each training set is discerned in subsequent steps of the algorithm. As presented in the **Results and Discussion**, the COREPA-C approach enables a quantitative evaluation of ATs and NATs by assessing similarity between APs and NAPs.

Step 2. Evaluation of stereoelectronic parameters hypothesized to be associated with biologically similar compounds. In the original version of COREPA [1], a restricted set of parameters, hypothesized to be associated with biological activity, were evaluated based on the normalized sum of dynamic similarity indices between each pair of molecules in a training set [2]. In COREPA-C, stereoelectronic parameters with maximal similarity among the chemicals within the training sets and with the least similarity to chemicals in different training sets (i.e., the most distinct APs and NAPs) are assumed to be most closely associated with the activity

under consideration. These parameters are used in the subsequent step of the algorithm.

Step 3. Recognition of the common reactivity pattern. The common reactivity patterns for active and nonactive chemicals are described in terms of the variation of descriptor x around its maximum probability value in the chemical distribution ($maxP_k^A(x)$ or $maxP_k^{NA}(x)$). The amount of variation around $maxP_k^A(x)$ and $maxP_k^{NA}(x)$ is used to define common reactivity patterns. The variation is set by the user in terms of confidence limits around selected pattern maxima; i.e., a specified portion of the distribution area around $maxP_k^A(x)$ or $maxP_k^{NA}(x)$.

The more narrow, or well defined, $P_k(x)$ will occur when the conformer distributions and chemicals are more similar (i.e., more in phase). A quantitative measure of common reactivity patterns can be attained by noting the amount of $P_k(x)$ variation at different confidence intervals. Better defined patterns are associated with a small increase in parameter variation, with increasing confidence levels around $maxP_k$, and higher values (i.e., greater intensity) of the associated probabilistic functions $P_k(x)$. Establishment of a similarity threshold between AP and NAP, i.e., S(AP/NAP), and cutoff thresholds is an empirical task, and needs to be validated by screening compounds not used in the training sets.

2.2 AR Ligands: Binding Affinity and Electronic Structure

The chemicals examined in this study are those AR ligands modeled in our description of the discrete COREPA algorithm [1]. Ligands #1–21 were used to derive training sets (Figure 2), while ligands #22–28 were used as a validation set (Figure 3). The AR binding affinities (pK_i) were obtained from Kelce *et al*. [13] and Waller *et al*. [14] and are based on a competitive binding assay using [^{3}H]RI881 (a radiolabeled synthetic androgen; see Kelce *et al*. [13]).

Conformers for each ligand were obtained from Mekenyan *et al*. [1], in which the 3DGEN procedure described by Ivanov *et al*. [12] was used to exhaustively generate energetically reasonable conformers. As in previous studies [1, 4, 5], conformers used in this analysis were restricted to those within 20 kcal/mol of the ΔH_f^o for the conformer associated with the absolute energy minimum, under the assumption that receptor binding or solvation effects could compensate conformer interconversion [11,15,16]. As noted in our previous study [1], the conformers of AR ligands within 20 kcal/mol of the lowest energy structure exhibited significant variation in electronic structure, which highlights the necessity of including all energetically-reasonable conformers when defining common reactivity patterns.

The global and local electronic descriptors used for the models was restricted to parameters hypothesized to be associated with AR binding affinity [11,14,17–19] and consistent with those used in our earlier study [1]. Stereoelectronic parameters were calculated with MOPAC 7 [20], augmented by a computing module that provided additional reactivity descriptors [21], using the AM1 all-valence electron, semi-empirical Hamiltonian (see the **Abbreviations** for the stereoelectronic parameters used).

3 Results and Discussion

To provide a point of reference, in our preceding application of the discrete version of COREPA, a reactivity pattern based on distances between electronegative atoms and their charges was found to discriminate active versus nonactive AR ligands [1]. An interatomic distance range of 10.2 to 11.1 Å with ten partitions of the distance range, derived for an AT of $pK_i \geqslant 0.7$ (a training set comprised of the six most active ligands) was identified as a less restrictive distance screen. This distance range was typically associated with O–O distances between the A- and D-rings (see Figure 1). A distance range of 10.7 to 11.1 Å, based on 20 partitions, was viewed as a more restrictive screen. Using the same training set (AT of $pK_i \geqslant 0.7$), the common range of atomic charges of −0.333 to −0.303 a.u., (analyzed with 20 partitions) was selected as a default screen, with a range of −0.333 and −0.313 a.u., based on 30 partitions, considered more restrictive. These reactivity patterns were validated by employing a one step screen that simultaneously addressed interatomic distance and atomic charge, and was designed to determine whether or not the algorithm could identify ligands with $pK_i \geqslant 0.7$. When employing the more restrictive screen, the algorithm was capable of discriminating "nonactive" ligands. The analysis indicated, however, that discrete partitioning of parameter distributions made it difficult to accurately quantify variation among conformer distributions. In addition, wide differences in conformational flexibility among some of the ligands made it difficult to interpret model predictions in certain instances.

3.1 Derivation of AR Ligand Reactivity Patterns

Training sets were established for active and nonactive ligands by defining AT and NAT, respectively, in terms of pK_i thresholds (Step 1). In the current study, ATs of 0.0, 0.7 and 1.0 with corresponding NATs of 0.0, −2.0 and −2.7 were used. Note that an AT of 0.7 and a NAT of −2.7 were used in our previous study [1].

The stereoelectronic descriptors that provided the greatest similarity among ligands within the training sets and the most distinct (nonoverlapping) reactivity patterns for "active" and "nonactive" ligands are assumed to be most

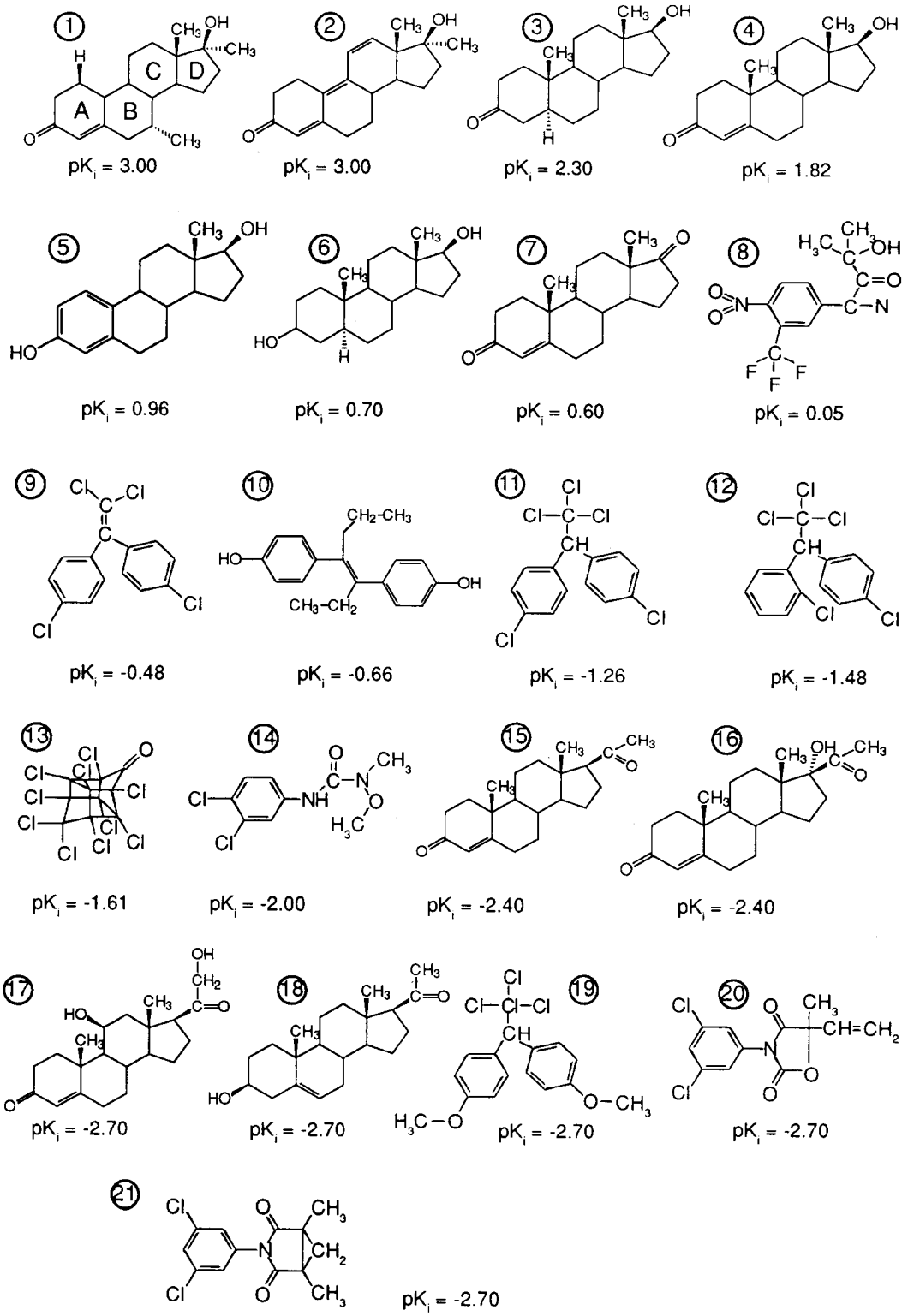

Figure 2. Structures of androgen receptor ligands used to establish training sets [pK_i values from 14]. Ligand names and the number of optimized conformers (n) with ΔH_f^o within 20 kcal/mol of the lowest energy conformer are as follows: #1) mibolerone (n = 9); #2) methylrienolone (n = 12); #3) 5α-dihydrotestosterone (n = 12); #4) testosterone (n = 5); #5) estradiol (n = 6); #6) 5α-androstane-3α, 17β-diol (n = 4); #7) Δ′-androstenedione (n = 5); #8) hydroxyflutamide (n = 19); #9) p,p′-DDE (n = 28); #10) diethylstibestrol (DES) (n = 126); #11) p,p′-DDT (n = 19); #12) o,p′-DDT (n = 24); #13) kepone (n = 1), #14) linuron (n = 26); #15 progesterone (n = 26); #16) 17α-hydroxyprogesterone (n = 22); #17) corticosterone (n = 35); #18) pregnolone (n = 13); #19) methoxyclor (n = 20); #20) vinclozolin (n = 11); #21) procymidone (n = 6). Rings A, B, C, and D identified in ligand #1. Ligands #1-6 represent the "active" ligands used to derive the common reactivity pattern for AR binding. Adapted from Mekenyan *et al.* [1].

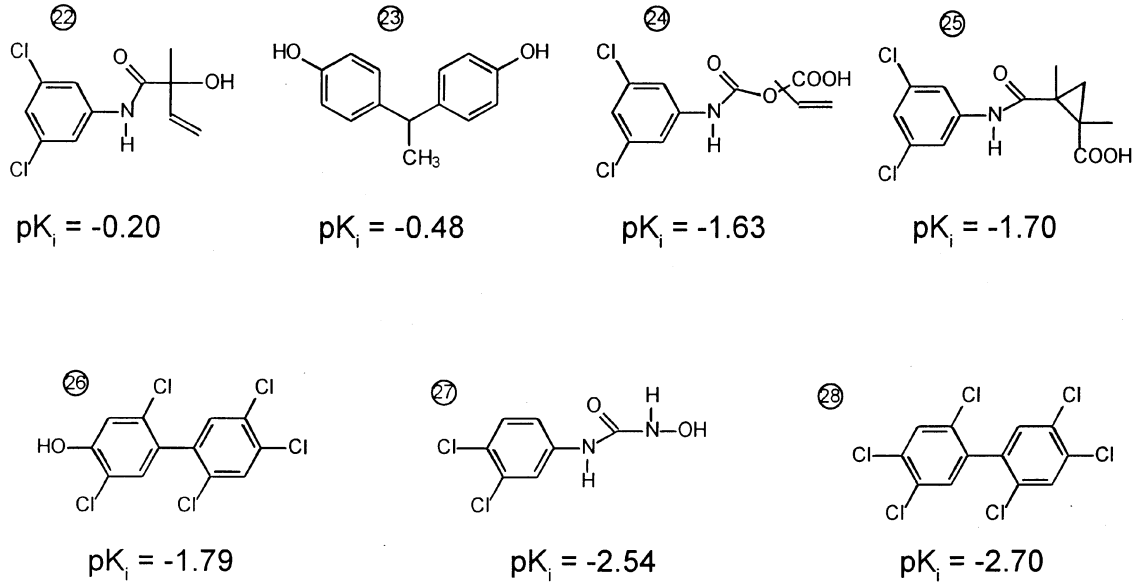

Figure 3. Structures of androgen receptor ligands used to establish validation sets [pK_i values from 14]. Ligand names are as follows: #22) 3,′5′-dichloro-2-hydroxy-2-methylbut-3-en anilide; #23) 2,2-bis(p-hydroxyphenyl)-1,1,1-trichloroethane; #24) 2-{[3,5-dichlorophenyl)carbamoyl]oxy}-2-methyl-3-butenoic acid; #25) 3,5-dichlorobenzanilide 2-cyclopropanecarboxylic acid; #26) the hydroxylated analog of PCB 153; #27) hydroxylinuron; #28) PCB 153. Adapted from Mekenyan *et al.* [1].

closely associated with the biological activity under consideration (Step 2). The similarity between chemicals in the training sets ("within group similarity") across various steric and electronic indices were calculated using Equation 7, at AT = 1.0, 0.7 and 0.0 and NAT = 0.0, −2.0 and −2.7. The calculated similarity measures are listed in Table 1. The requirement for high "within group similarity" is especially significant for the training sets of active chemicals. Within group similarity for the nonactive training set is less of an issue because of the large diversity of chemicals present in this particular data set. Similarity between reactivity patterns of active and nonactive chemicals, calculated for different pairs of ATs and NATs, also are presented in Table 1. Results in Table 1 show that local molecular descriptors, such as distances between heteroatoms (R), d(R_R), and charges of oxygens, Q(O), provide high within group similarity, but relatively low between group similarity. High within group similarity for the training sets of active ligands also were observed for acceptor and donor delocalizabilities, SN(R) and SE(R), respectively; however, both parameters demonstrated high similarity between patterns of active and nonactive ligands.

The discrimination between the training sets of active and nonactive ligands, provided by d(R_R) and Q(R), is further illustrated in Figure 4. As depicted in Figures 4a and 4b, d(R_R) provided greater distinction between active and passive patterns for all heteroatoms, whereas Q(R) was more discriminatory when R was restricted to oxygen only (Figure 4c and 4d). A similar observation was observed with SN(R), and SE(R) (data not shown). Based on these results, d(R_R) and Q(O) were incorporated in the third step of the analysis.

As noted previously, a pK_i of 0.7 was empirically defined as the active threshold in our study using the discrete version of the COREPA algorithm [1]. The COREPA-C algorithm, however, provides the means to quantitatively describe the variation of chemical similarity for different training sets. The analysis of similarity within and between a variety of pairings of APs and NAPs, with respect to d(R_R) and Q(O), suggested that the greatest discrimination between patterns and similarity within the APs was associated with an AT of 0.7 and a NAT of −2.0 (see Table 1).

A leave-one-out analysis also indicated that the AP derived with an AT of 0.7 (ligands #1–6) was more stable than an AP derived with an AT of 0.0 (ligands #1–8). For example, with an AT of 0.7, the 50% confidence limit for d(R_R) was 10.38 to 11.07 Å ($\Gamma = 3$). The mean distance range from the leave-one-out analysis was 10.34 (10.30–10.39) to 11.10 (11.06–11.14) Å. The d(R_R) with an AT of 0.0 was 10.27 to 10.94 Å. In addition to a shift to a smaller interatomic distance range with the inclusion of ligands #7 and 8, there was also an increase in the variability of the pattern based on a leave-one-out analysis. The mean distance range from the leave-one-out analysis with an AT of 0.0 was 10.24 (10.17–10.37) to 10.94 (10.90–11.01) Å. The stability of APs can also be assessed in terms of D. The mean D values for APs derived from leave-one-out analyses for ligands #1–6 compared to the AP for the intact training set were 0.487 and 0.0058, respectively, for Q(O) ($\Gamma = 0.03$) and d(R_R)

Table 1. Within-group average similarity ($S_{yz}(x)$) and similarity between active and nonactive patterns (S(AP/NAP)) for continuous distributions of stereoelectronic parameters approximated using default values of Γ. Active (ATs) and nonactive thresholds (NATs) based on pK_i values of 1.0, 0.7 and 0.0 and 0.0, −2.0 and −2.7, respectively.

		Average $S_{yz}(x)$ [%]						S(AP/NAP) [%]		
Parameters[a]	Γ	AT = 1.0	AT = 0.7	AT = 0.0	NAT = 0.0	NAT = − 2.0	NAT = − 2.7	AT = 1.0 NAT = 0.0	AT = 0.7 NAT = − 2.0	AT = 0.0 NAT = 2.7
Planarity	5.453	0.00	0.00	0.00	29.98	21.40	29.92	18.45	27.47	68.94
Max Distance	1.675	0.00	0.00	0.00	27.30	16.63	20.71	13.19	22.29	75.25
Geom. Wiener	735.227	0.00	1.77	1.89	35.9	21.43	30.00	1.03	1.08	16.26
Electronegativity	0.365	69.08	42.67	37.86	36.41	39.91	33.30	41.37	37.90	82.16
Vol.P.	0.095	63.89	59.67	49.40	7.52	6.76	6.59	0.02	2.34	3.83
E_{HOMO}	0.295	34.99	33.87	41.58	24.73	28.76	27.26	11.83	36.25	10.50
E_{LUMO}	0.641	47.34	36.89	36.68	60.49	61.11	52.04	37.81	35.68	48.15
E-gap	0.702	34.06	26.30	32.60	56.93	56.28	51.82	12.30	75.98	52.28
Dipole Moment	1.174	60.08	49.72	46.97	64.85	90.15	89.62	41.35	43.99	63.59
Distance -all	1.629	98.49	98.76	98.42	99.55	100.0	100.0	73.77	93.84	84.42
Charge - all	0.119	96.57	94.60	90.29	16.94	16.72	16.75	56.93	92.57	80.71
Donor Deloc - all	0.030	92.37	82.46	78.80	4.33	4.35	4.36	64.73	95.90	77.97
Acceptor Deloc - all	0.042	90.35	83.86	76.63	6.02	6.07	6.08	11.69	67.19	43.25
d(R-R)	1.629	90.66	90.78	72.67	70.98	70.90	75.40	1.26	9.39	1.12
d(O-O)	1.629	90.66	90.78	75.39	25.92	58.71	57.14	79.19	36.23	13.01
Q(R)	0.119	99.34	92.81	86.33	14.66	15.35	15.33	39.70	65.24	53.31
Q(O)	0.028	95.04	79.24	70.98	2.22	4.05	4.06	25.62	16.40	11.07
SE(R)	0.028	85.29	62.05	60.27	3.90	3.96	3.97	14.05	63.97	44.95
SE(O)	0.010	68.01	42.13	42.95	0.88	1.55	1.55	5.95	37.42	14.18
SN(R)	0.029	95.57	93.99	83.2	4.05	4.07	4.06	63.51	81.67	68.27
SN(O)	0.008	82.48	73.33	63.25	0.66	1.16	1.15	38	26.47	5.7
QH(R)	0.147	65.45	67.32	66.13	21.13	21.26	21.34	76.19	78.67	71.68
QH(O)	0.138	64.34	66.35	64.29	10.09	19.72	19.16	89.02	83.34	70.00
QL(R)	0.139	69.70	67.46	69.61	18.98	18.78	17.61	86.54	95.64	77.67
QL(O)	0.139	69.70	67.46	67.59	9.86	18.83	18.30	87.54	95.16	76.78
Pol(R)	0.004	91.86	88.72	78.25	0.54	0.54	0.54	52.49	59.49	57.93
Pol(O)	0.001	83.30	77.23	72.34	0.12	0.20	0.20	74.04	40.12	30.00

[a] See **List of Abbreviations** for explanation of parameters.

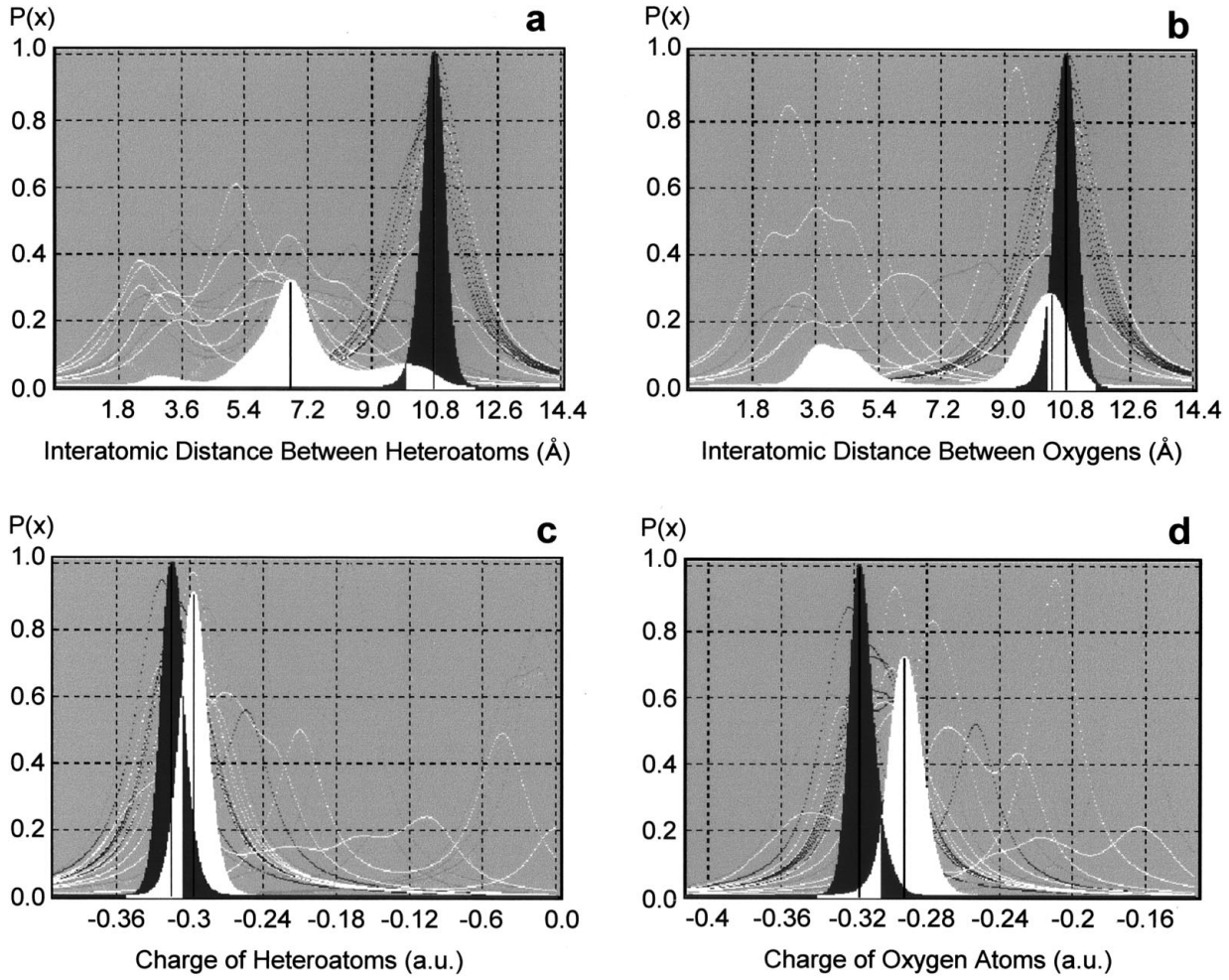

Figure 4. Reactivity patterns of active ($pK_i \geqslant 0.7$) and nonactive ($pK_i \leqslant -2.0$) androgen receptor ligands based on a) interatomic distances between heteroatoms; b) interatomic distances between oxygen atoms only; c) charge on heteroatoms; and d) charge on oxygens only. Interatomic and charge distributions were approximated using Γ values of 1.629 and 0.028, respectively. The common distribution curves for active chemicals are red, while the distribution for nonactive chemicals is white. Distributions from ligands with pK_i values between 0.7 and -2.0 are colored in green.

($\Gamma = 3.0$). An analysis based on ligands #1–8 resulted in mean D values of 0.769 and 0.0083 for Q(O) and d(R_R) respectively. Thus, the COREPA-C analysis supported the empirical conclusion [1] that a pK_i threshold of 0.7 was reliable for distinguishing AR ligand binding affinity for this data set.

Using a pK_i of 0.7 as the AT, reactivity patterns based on characteristics of d(R_R) and Q(O) were derived (Step 3). Variation of distance and charge ranges, associated with 10 and 50% confidence limits around $\max P_k^A(x)$, were assessed for Γ values of 1 to 9 and 0.01 to 0.1, respectively (see Tables 2 and 3). To assess the generality of reactivity patterns for d(R_R) and Q(O), S(AP/NAP) and Cut-off(AP/NAP) ranges can be evaluated in terms of varying Γ values. These data illustrate how the use of increasing Γ values to approximate continuous distributions lead to wider parameter ranges for d(R_R) and Q(O) in reactivity patterns (presented in terms of 10 and 50% confidence limits in Tables 2 and 3). With increasing Γ values, wider ranges of d(R_R) were associated with three- to five-fold increases in S(AP/NAP) and Cut-off(AP/NAP), while wider ranges of Q(O) were associated with 50- to 100-fold increases in S(AP/NAP) and Cut-off(AP/NAP). The increases in S(AP/NAP) and Cut-off(AP/NAP) provide relative measures of the increasing probability of incorrectly identifying a nonactive ligand as active (i.e., incorrectly predicting a $pK_i \geqslant 0.7$). It is important to note that even with the most restrictive d(R_R) and Q(O) ranges listed in Tables 2 and 3, there is overlap between the APs and NAPs and nonzero percentages of the NAPs are exceeded by maximum values associated with the AP. As compounds become more stereoelectronically similar to ligands #1–6 in terms of d(R_R) and d(O) (i.e., increasing percentages of compounds' distributions overlapping with the AP) their pK_i values should approach or exceed 0.7.

The effect of conformer selection on reactivity patterns was investigated by analyzing variation of S(AP/NAP) and parameter ranges associated with the threshold for $\Delta\Delta H_f^o$, as well as by employing the Boltzman statistic to weight conformers. Reduction of $\Delta\Delta H_f^o$ thresholds from 20 to 1 kcal/mol increased S(AP/NAP) from 18% to 21% and from 13% to 15% for reactivity patterns across Q(O) and d(R_R), respectively (Γ of 0.03 and 3). The increase in overlap between the patterns could be related to the reduced number of conformers in the training sets. The similarity between overall reactivity patterns of active chemicals for $\Delta\Delta H_f^o = 20$ and 1 kcal/mol was also assessed by D and S(K/AP), with $D = 0.57$ and $S = 92\%$, and $D = 0.06$ and $S = 85\%$, for Q(O) and d(R_R) based patterns, respectively ($\Gamma = 0.03$ and 3). Analogously, no significant variation of Q(R) ranges was observed with the decrease of $\Delta\Delta H_f^o$; however, d(R_R) ranges shifted from 10.38 to 11.07 Å to 10.57 to 10.92 Å. Similarly, the Boltzman weighting scheme resulted in slight variations in activity patterns. For example, using this weighting scheme a range of 10.45 Å to 11.12 Å was associated with the d(R_R) pattern, ($\Gamma = 3$; 50% confidence limit), compared to a range of 10.38 to 11.07 Å without weighting.

An examination of Tables 2 and 3 also illustrates how the combination of Γ value and confidence limit selection provides flexibility in defining the specificity of reactivity patterns. Using a 50% confidence limit, the d(R_R) range of 10.68 to 10.95 Å increased to 9.79 to 11.57 Å as Γ increased from 1 to 9, while with a 10% confidence limit the d(R_R) range increased from 10.79 to 10.84 Å to 10.51 to 10.85 Å. In terms of Q(O), the 10% and 50% confidence limit ranges were −0.318 a.u. and −0.320 to −0.317 a.u., respectively, with a Γ of 0.01. With a Γ of 0.09, the 10 and 50% confidence limits were −0.314 to −0.310 a.u. and −0.323 to −0.300, respectively.

The ability of different reactivity patterns to screen compounds for AR binding affinity was evaluated by analyzing ligands #9–21(i.e., ligands with pK_i values $\leqslant -0.48$; see Figure 2), which were represented by 202 conformers. Though charge-based patterns were derived from Q(O), the patterns were applied against wildcard heteroatoms to illustrate the ability of the algorithm to assess similarity without the need to predetermine a toxicophore or establish an alignment against a lead, or template, molecule. Thus, the interatomic distances and charges were not specified to hydroxyl or carbonyl oxygens, nor was an automated or manual assignment of A- and/or D-ring steroidal counterparts required. All active patterns derived from Tables 2 and 3, described below, were combined with an additional requirement that active ligands could not have a d(R_R) between 2.0 to 9.0 Å. This restriction was based on the observation that at several Γ values all nonactive ligands (i.e., pK_i values $\leqslant -2.0$) had interatomic distances within this range, while active ligands (i.e., pK_i values $\geqslant 0.7$) did not (e.g., see Figure 4).

Table 2. For interatomic distances between heteroatoms, similarity between reactivity patterns (S(AP/NAP)), the portion of nonactive patterns exceeded by the maximum of the active pattern (Cut-off(AP/NAP)) (based on 50% confidence levels), and the 10 and 50% confidence levels around the maximum interatomic distance probability based on continuous distributions derived from ligands with pK_i values $\geqslant 0.7$ and $\leqslant -2.0$. Distributions approximated with values of Γ between 1 and 9.

Γ	S(AP/NAP) [%]	Cut-off (AP/NAP) [%]	10% C.I. [Å]	50% C.I. [Å]
1	6.60	1.62	10.79–10.84	10.68–10.95
2	11.27	2.00	10.71–10.80	10.51–11.00
3	13.88	2.07	10.66–10.79	10.38–11.07
4	16.67	2.13	10.63–10.79	10.27–11.15
5	19.77	2.47	10.60–10.80	10.15–11.24
6	23.03	2.78	10.57–10.81	10.05–11.33
7	26.32	3.20	10.55–10.82	9.96–11.41
8	29.58	3.69	10.53–10.84	9.87–11.50
9	32.76	4.25	10.51–10.85	9.79–11.57

Table 3. For charges on oxygen atoms, similarity between reactivity patterns (S(AP/NAP)), the portion of nonactive patterns exceeded by the maximum of the active pattern (Cut-off(AP/NAP)) (based on 50% confidence levels), and the 10 and 50% confidence levels around the maximum charge probability based on continuous distributions derived from ligands with pK_i values $\geqslant .7$ and $\leqslant -2.0$. Distributions approximated with values of Γ between 0.01 and 0.09.

Γ	S(AP/NAP) [%]	Cut-off (AP/NAP) [%]	10% C.I. [a.u]	50% C.I. [a.u]
0.01	1.11	0.09	− 0.318	− 0.320 to −0.317
0.02	8.65	0.30	− 0.318 to −0.317	− 0.320 to −0.315
0.03	17.93	0.81	− 0.318 to −0.316	− 0.322 to −0.312
0.04	26.20	1.93	− 0.317 to −0.315	− 0.322 to −0.310
0.05	33.17	3.67	− 0.316 to −0.313	− 0.321 to −0.307
0.06	38.88	5.30	− 0.315 to −0.311	− 0.321 to −0.305
0.07	43.52	7.01	− 0.314 to −0.311	− 0.322 to −0.304
0.08	47.30	8.02	− 0.314 to −0.310	− 0.322 to −0.302
0.09	50.44	9.66	− 0.314 to −0.310	− 0.323 to −0.300

As discussed previously, distributions based on larger Γ values and the use of larger confidence limits will lead to wider ranges in d(R_R) and Q(O). In general, increasing parameter ranges are associated with a greater likelihood that ligands will be incorrectly identified as having a pK_i value > 0.7; i.e., an increasing rate of false positive identifications. Smaller parameter ranges increase the rate of false negative identifications; i.e., ligands incorrectly identified as having a pK_i value < 0.7. It is important to note that due to the probabilistic nature of the algorithm, it is possible that the choice of an active pattern based on a small Γ or confidence limit can lead to incorrect classifications of conformers from an active ligand. For example, using d(R_R) in a single parameter screen of 10.68 to 10.95 Å ($\Gamma = 1.0$; 50% confidence limit; see Table 2), all of the conformers of ligand #6 were incorrectly identified as having a pK_i value < 0.7. If Q(R) is used in a single parameter screen of -0.320 to -0.317 a.u. ($\Gamma = 0.01$; 50% confidence limit; see Table 3) all of the conformers of ligand #1 were incorrectly identified as having a pK_i value < 0.7. Consequently larger Γ values were required to establish patterns that would include all the conformers from all of the active ligands. These single parameter screens do, however, lead to "false positive" classifications of nonactive conformers. For example, a d(R_R)-based pattern of 10.38 to 11.07 Å ($\Gamma = 3.0$; 50% confidence limit) incorrectly identified 3, 10, 20, 19 and 7 conformers of ligands #7, #15, #16, #17 and #18 as having pK_i values $\geqslant 0.7$. The similarity of ligand distributions for d(R_R) to the corresponding AP ranged from 5.95% for ligand #17 to 17.4% for ligand #18.

Ultimately, a two-parameter screen based on d(R_R) and Q(R) was used to minimize the rate of false positive identifications. Thus, for a ligand to be classified as active at least one conformer was required to fall within the d(R_R) and Q(R) APs. A screening pattern with a d(R_R) of 10.38 to 11.05 Å ($\Gamma = 3.0$; 50% confidence limit, see Table 2) combined with a Q(R) of -0.322 to -0.312 a.u. ($\Gamma - 0.03$; 50% confidence limit; see Table 3) correctly predicted that the pK_i values of ligands #7–21 are < 0.7 (i.e., no conformers from these ligands fell within both APs). These ranges were representative of the least restrictive reactivity patterns that correctly discriminated the ligands. Larger parameter ranges increased the incidence of false positive identifications. For example, with a d(R_R) of 10.38 to 11.07 Å ($\Gamma = 3$; 50% confidence limit) and a Q(R) of -0.323 to -0.300 a.u. ($\Gamma = 0.09$; 50% confidence limit), ten conformers of ligands #15, 16 and 18 were incorrectly identified as having pK_i values > 0.7. Reducing Q(R) to -0.322 to -0.302 ($\Gamma = 0.08$; 50% confidence limit) one conformer of ligand #15 was incorrectly identified; reducing the range to -0.322 to -0.304 a.u. lead to the correct classification of ligands #9–21. Holding Q(R) at -0.322 to -0.300, but decreasing the d(R_R) range to 10.51 to 11.00 Å ($\Gamma = 2$; 50% confidence limit), four conformers of ligands #15, 16 and 18 were incorrectly classified as having a pK_i value of > 0.7. Further decreasing the d(R_R) range to 10.68 to 10.95 Å ($\Gamma = 1$; 50% confidence limit) resulted in two conformers of ligands #16 and 18 being incorrectly identified and all of the conformers of ligand #6 being incorrectly identified as having pK_i values < 0.7.

As described above, the classification of an "unknown" ligand was based on whether or not one or more conformers for the compound fell within both the d(R_R) and Q(R) APs. Alternatively, values of D or S(k/AP) for comparisons of distributions of unknown compounds to the APs could be used to screen ligands. For the d(R_R) AP, D and S(k/AP) values for ligands #1-6 varied from 0.135 to 0.144 and 16.7 to 18.9%, respectively. For the Q(R) AP, the D and S(k/AP) values ranged from 8.87 to 10.67 and 11.2 to 18.2%. Using S(k/AP) as a criteria, thresholds of 16.7 and 11.2% for d(R_R) and Q(R) could be used in a two-parameter screen. This screen identified ligand #18 (S(k/APs) for d(R_R) and Q(R) of 17.4 and 11.8%) as a potentially active ligand. For the remaining ligands, S(k/AP) values for d(R_R) and Q(R) ranged from 17.3 to 1.2% and from 12.4 to 1.1%, respectively. Using D or S(k/AP) as criteria to assess unknown ligands facilitates a more quantifiable approach for determining similarity in reactivity patterns. The use of these similarity metrics also permits the sensitivity of different thresholds to be more readily evaluated in terms of potential rates of false negative and positive classifications of unknown ligands.

The reactivity pattern of 10.38 to 11.07 Å, for d(R_R), and -0.322 to -0.312 a.u., for Q(R), was assessed against an external validation set comprised of the seven compounds in Figure 3. As noted above, this reactivity pattern was the least restrictive that correctly discriminated conformers from ligands #7–21. The conformer generation routine and subsequent quantum chemical optimization produced a total of 132 conformers for the seven compounds (all within 20 kcal/mol of the lowest energy geometries; [1]). All seven of these compounds were properly discriminated as ligands likely to exhibit a $pK_i < 0.7$. However, these results should be viewed with caution because only a small number of compounds were available across pK_i ranges associated with potential ATs. Research in progress with estrogen receptor ligands, in which a greater number of compounds with selected binding affinity ranges are available, will permit a more detailed evaluation of reactivity patterns derived with the COREPA-C algorithm.

The present validation tests were based on reactivity patterns derived from two descriptors. The discrimination ability of the algorithm could conceivably be further improved by making use of reactivity patterns based on other molecular descriptors shown to be associated with receptor binding affinities (e.g., delocalizabilities; data not shown). Given the

relatively small number of ligands in this data set, use of additional descriptors would, of course, be of questionable significance. However, the use of interatomic distance and atomic charge to identify AR ligands with different levels of binding affinity is mechanistically reasonable, and as described previously [5], consistent with general concepts concerning the nature of steroid hormone ligand binding sites [18,19,22,23].

3.2 Reactivity Patterns and Toxicophore Sites

The patterns derived with COREPA-C also provide insights as to the toxicophore sites of AR ligands, which was not readily attained with the discrete COREPA algorithm. Using a pK_i of 1.0 as an AT and $\Gamma \leqslant 0.02$, two maxima were distinguished in the charge based pattern and correspond to $\max P_k^A(\ Q(O))$ of -0.302 a.u. and -0.318 a.u. (Figure 5a). The analysis showed that the maximum at -0.318 a.u. corresponds to 17 β-OH, whereas the maximum at -0.302 a.u. is associated with the α,β-unsaturated ketone at position 3. Thus, with a more precise description of the distribution (smaller Γ) and a higher active threshold (more precise estimate of biological similarity) the algorithm could distinguish two different types of electronegative atoms in active androgen ligands.

An examination of Figure 5 also indicates that the difference between the AP and NAP is mainly due to the properties of the 17 β-OH. In active androgen ligands, this oxygen appears to be more negatively charged than the carbonyl oxygens at position 3. Moreover, the corresponding conformer distributions across acceptor- and donor-delocalizabilities (Figures 5b and 5c) indicate that the carbonyl oxygens at position 3 have higher electron donor and acceptor properties than the 17-β hydroxyl oxygens in active androgens, even though higher electron density is localized at this oxygen. The distribution of atom polarizabilities also indicated that this charge is more localized than the charge on carbonyl oxygens at position 3 (data not shown). Thus, in active androgens, the 17-β oxygens are less reactive than the oxygens at position 3, due to lower charge delocalizability. There is, however, no difference in reactivity between 17-β oxygens and those at position 3 in nonactive ligands, which explains the presence of a single maximum for these sites in the NAPs. In conclusion, it appears that the higher pK_i of active androgen ligands is due, in part, to lower reactivity of electronegative sites in the 17-β position compared to the nonactive ligands.

4 Summary and Conclusions

In the COREPA-C algorithm the continuous approximation of conformer distributions provides a probabilistic interpretation of reactivity patterns. This advancement addresses several limitations of the original COREPA algorithm [1]. These limitations concern the ability to adequately resolve the influence of parameter partitioning on defining reactivity patterns, biased 3-D similarity assessments between molecules with large differences in conformational flexibility, and the inability to quantitatively assess similarity between conformer distributions of individual chemicals and common reactivity patterns. The first of these limitations is addressed by assessing the precision of the continuous approximation (i.e., the magnitude of the Γ parameter) on reactivity patterns. The second limitation is resolved by normalizing conformer distributions for each ligand, and using similarity indices based on the intersection of ligand distributions. Finally, the probabilistic character of conformer distributions allows comparison of the distributions of individual molecules to the distributions of training sets as well as comparisons between training sets. This capability permits an assessment of user selected activity thresholds for defining training sets and evaluation of the local electronic character of toxicophores. It should be noted that while the COREPA-C alogrithm allows the user to assess chemical similarity in terms of two or more descriptors simultaneously, the algorithm is limited in the sense that interdependencies of the conformer distributions are not handled in a strict multivariate manner.

Through the selection of Γ and/or confidence limits around distribution maxima, COREPA-C also provides flexibility in developing screening algorithms for different hazard identification needs. This feature of COREPA-C is especially significant when contemplating the application of the approach for large data bases of 3-D structures, where it is likely that only a single conformer per compound will be available. In these situations, we envision the use of an initial, less restrictive, screening strategy based on a comparatively wide confidence interval and/or large Γ values. This screen would be designed to minimize the percentage of false negatives (i.e., compounds incorrectly predicted to be below a specified activity threshold). A subsequent screen of the resulting "positive" compounds, using a multiple conformational analysis, could be invoked based on a more restrictive pattern to eliminate "false positives." The application of this approach for larger data sets is in progress.

Acknowledgments

This research was supported, in part, by a U.S. EPA Cooperative Agreement (CR822306-01-0) with the Bourgas University "As. Zlatarov", as well as by Contract No. 12231-96-10 F1ED ISP BG between Bourgas University and the Environment Institute of the European Union. The authors thank Dr. Gilman Veith and Prof. Gerrit Schüürmann for valuable discussions, and Dr. Julian Ivanov for his active involvement in developing the recent version of the method. We also thank Dr. Mumtaz Pasha and Ms. Christine Russom

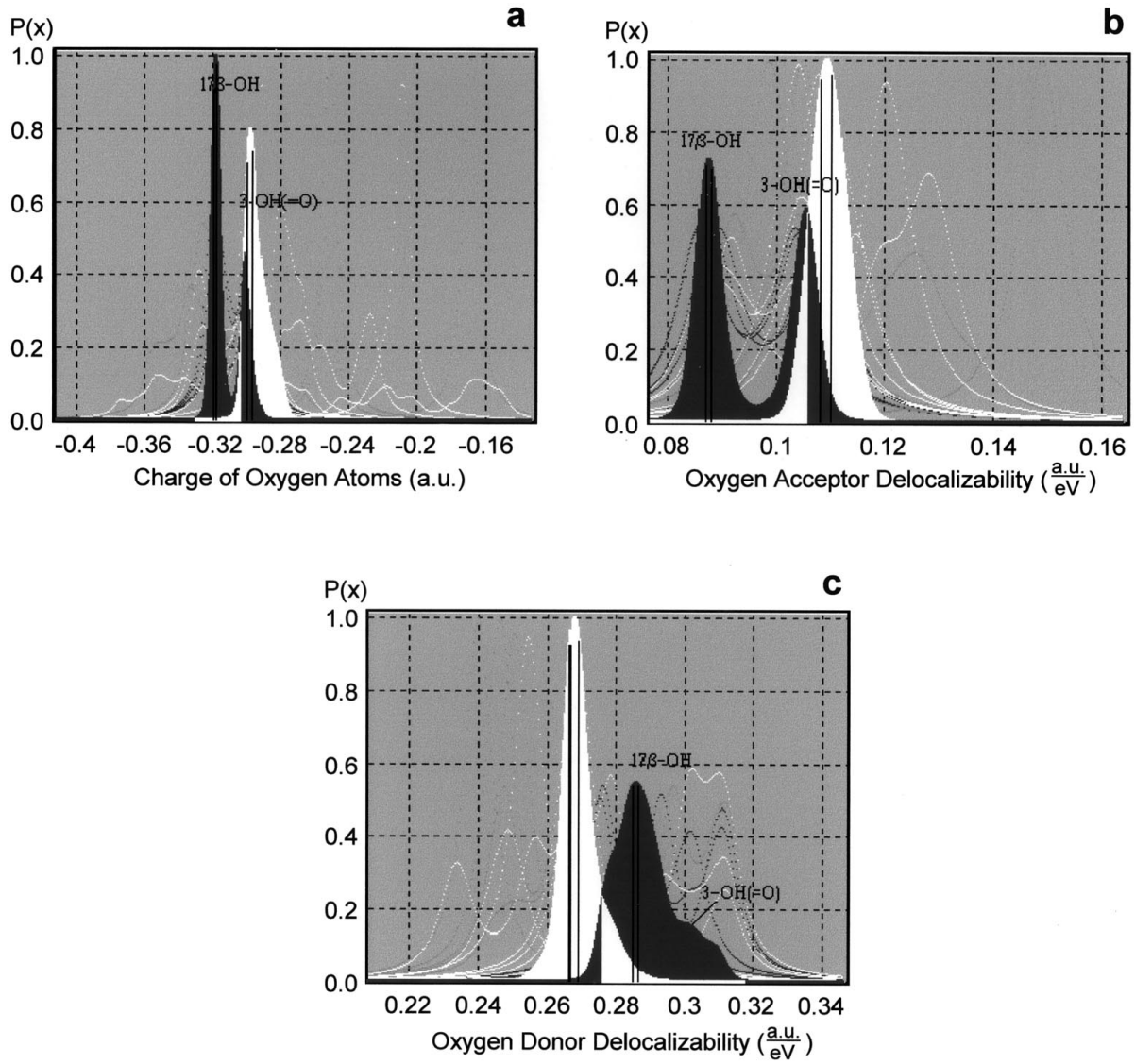

Figure 5. Reactivity patterns of active ($pK_i \geqslant 1.0$) and nonactive ($pK_i \leqslant -2.0$) androgen receptor ligands based on: a) charge on oxygen; b) oxygen acceptor delocalizabilities; and c) oxygen donor delocalizabilities. Distributions were approximated using a Γ value of 0.01.

for valuable review comments. Roger LePage and Diane Spehar provided assistance in manuscript preparation. The work has been accomplished in the framework of a collaborative agreement between the U.S. EPA and European Chemicals Bureau at the Environment Institute of the European Union.

Disclaimer

Mention of models or modeling approaches does not constitute endorsement on the part of the U.S. EPA.

References

[1] Mekenyan, O.G., Ivanov, J.M., Karabunarliev, S.H., Bradbury, S.P., Ankley, G.T. and Karcher W., COREPA: A new approach for the elucidation of COmmon REactivity PAtterns of chemicals. I. Stereoelectronic requirements for androgen receptor binding. *Environ. Sci. Technol. 31*, 3702–3711 (1997).

[2] Mekenyan, O.G., Ivanov, J.M., Karabunarliev, S.H., Hansen, B., Ankley, G.T. and Bradbury SP., A new approach for estimating 3-D similarity that incorporates molecular flexibility. in: Chen, F. and Schrmann, G., (Eds.), *Quantitative Structure Activity Relationships in Environmental Sciences-VII*, SETAC, Pensacola, FL, 1998, pp. 39–57.

[3] Eliel, E.L., Chemistry in three dimensions. in: Warr, W.A. (Ed.), *Chemical Structures*, Vol. 1, Springer, Berlin, Germany, 1993, pp 1–8.

[4] Bradbury, S.P., Mekenyan, O.G. and Ankley, G.T., Quantitative structure-activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity: An assessment of conformational flexibility. *Environ. Chem. Toxicol. 15*, 1945–1954 (1996).

[5] Bradbury, S.P., Mekenyan, O.G. and Ankley, G.T., The role of ligand flexibility in predicting biological activity: Structure-activity relationships for aryl hydrocarbon, estrogen and

androgen receptor binding affinity. *Environ. Toxicol. Chem 17*, 15–25 (1998).
[6] Mekenyan, O.G., Ivanov, J.M., Veith, G,D and Bradbury, S.P., DYNAMIC QSAR: A new search for active conformations and significant stereoelectronic indices. *Quant. Struct.-Act. Relat. 13*, 302–307 (1994).
[7] Mekenyan, O.G., Schultz, T.W., Veith, G.D. and Kamenska, V.B., "Dynamic" QSAR for semicarbazide-induced mortality in frog embryo. *J. Appl. Toxicol. 16*, 355–363 (1996).
[8] Mekenyan, O.G., Veith, G.D., Call, D.J. and Ankley, G.T., A QSAR evaluation of Ah receptor binding of halogenated aromatic xenobiotics. *Environ. Health Perspect. 104*, 1302–1309 (1996).
[9] Prendergast, K., Adams, K., Greenlee, W.J., Nachbar, R.B., Patchett, A.A. and Underwood, D.J., Derivation of a 3D pharmacophore model for the angiotensin-II site one receptor. *J. Comput.-Aided Mol. Des. 8*, 491–512 (1994).
[10] Veith, G.D., Mekenyan, O.G., Ankley, G.T. and Call D.J., QSAR evaluation of α-terthienyl phototoxicity. *Environ. Sci. Technol. 29*, 1267–1272 (1995).
[11] Wiese, T. and Brooks, S.C., Molecular modeling of steroidal estrogens: Novel conformations and their role in biological activity. *J. Steroid Biochem. Mol. Biol. 50*, 61–72 (1994).
[12] Ivanov, J.M., Mekenyan, O.G., Bradbury, S.P. and Schuurmann, G., A kinetic analysis of the conformational flexibility of steroid hormones. *Quant. Struct.-Act. Relat. 17*, 437–449 (1998).
[13] Kelce, W.R., Monosson, E., Gamasik, M.P., Laws, S.C. and Earl Gray, L., Jr., Environmental hormone disruptors: Evidence that vinclozolin developmental toxicity is mediated by antiandrogenic metabolites. *Toxicol. Appl. Pharmacol. 126*, 276–285 (1994).
[14] Waller, C.L., Juma, B.W., Earl Gray, L., Jr. and Kelce, W.R., Three-dimensional quantitative relationships for androgen receptor ligands. *Toxicol. Appl. Pharmacol. 137*, 219–227 (1996).
[15] Anstead, G.M., Wilson, S.R. and Katzenellebogen, J.A., 2-Arylindenes and 2-arilindenones: Molecular structures and considerations in the binding orientation of unsymmetrical non-steroidal ligands to the estrogen receptor. *J. Med. Chem. 32*, 2163–2171 (1989).
[16] Ivanov, J.M., Karabunarliev, S.H. and Mekenyan, O.G., 3DGEN: A system for an exhaustive 3D molecular design. *J. Chem. Inf. Comput. Sci. 34*, 234–243 (1994).
[17] Waller, C.L., Minor, D.L. and McKinney, J.D., Examination of the estrogen receptor binding affinities of polychlorinated hydroxybiphenyls using three-dimensional quantitative structure-activity relationships. *Environ. Health Perspect. 103*, 702–707 (1995).
[18] VanderKuur, J.A, Wiese, T. and Brooks, S.C., Influence of estrogen structure on nuclear binding and progesterone receptor induction by the receptor complex. *Biochemistry 32*, 7002–7008 (1993).
[19] Wurtz, J.M., Bourguet, W., Renaud, J.P., Vivat, V., Chambon, P., Moras, D. and Gronemeyer H., A canonical structure for the ligand-binding domain of nuclear receptors. *Nature Struct. Biol. 3*, 87–94 (1996).
[20] Stewart, J.J.P., MOPAC: A general molecular orbital packages; version 7.0. software. Quantum Chemistry Program Exchange No. 455, University of Indiana, Bloomington, IN (1994).
[21] Mekenyan, O.G., Karabunarliev, S.H., Ivanov, J.M., and Dimitrov, D.N., A New development of the OASIS computer system. *Comput. & Chem. 18*, 173–187 (1994).
[22] Goldstein, R.A., Katzenellenbogen, J.A., Luthey-Schulten, Z.A., Seielstad, D.A. and Wolynes, P.G., Three-dimensional model for the hormone binding domains of steroid receptors. *Proc. Natl. Acad. Sci. USA (Biochemistry) 90*, 9949–9953 (1993).
[23] Lewis, D.F.V., Parker, M.G. and King, R.J.B., Molecular modeling of the human estrogen receptor ligand interactions based on site-directed mutagenesis and amino acid sequence homology. *J. Steroid Biochem. Mol. Biol. 52*, 55–65 (1995).

Received January 7, 1999; accepted on March 22, 1999