1997

A Computationally-Based Hazard Identification Algorithm That Incorporates Ligand Flexibility. 1. Identification of Potential Androgen Receptor Ligands

Steven P. Bradbury
Ovanes Mekenyan
Julian Ivanov
Stoyan Karabunarliev
Gerald T Ankley, et al.

A Computationally-Based Hazard Identification Algorithm That Incorporates Ligand Flexibility. 1. Identification of Potential Androgen Receptor Ligands

OVANES MEKENYAN,
JULIAN IVANOV, AND
STOYAN KARABUNARLIEV

Bourgas University "As. Zlatarov", 8010 Bourgas, Bulgaria

STEVEN P. BRADBURY,* AND
GERALD T. ANKLEY

U.S. Environmental Protection Agency, National Health and Environmental Effects Research Laboratory, Mid-Continent Ecology Division, 6201 Congdon Boulevard, Duluth, Minnesota 55804

WALTER KARCHER

Joint Research Centre, Environment Institute, European Chemicals Bureau, I-21020 Ispra (Varese), Italy

To advance techniques for screening large data sets of diverse structures for toxicologically active compounds, an algorithm was developed that is not dependent upon a predetermined and specified toxicophore or an alignment of conformers to a lead compound. Instead, the approach provides the means to identify and quantify specific global and local stereoelectronic characteristics associated with active compounds through a comparison of energetically-reasonable conformer distributions for specific descriptors. To illustrate the algorithm, the stereoelectronic requirements associated with the binding affinity of 28 steroidal and non-steroidal ligands to the androgen receptor were defined. Common ranges of interatomic distances, atomic charges, and atom polarizabilities of oxygen atoms for conformers of the ligands with the highest affinity for the androgen receptor (most active) did not overlap with those identified for conformers with the lowest binding affinity (least active). Using a set of stereoelectronic parameters that provided a maximal measure of pair-wise similarity among the conformers of the most active ligands, a model was developed to screen compounds for binding affinity. The model was capable of discriminating inactive ligands, as defined by a specified binding affinity threshold. This modeling technique could be a useful initial component in an integrated approach of employing computational and toxicological techniques in hazard identifications for large databases.

Introduction

Developing structure−activity relationships to screen large data sets of diverse chemical structures for toxicological activity in a technically-sound manner is challenging. Models capable of identifying relevant molecular characteristics that result in similar levels of activity require a clear definition of the toxicological end point(s) of concern to define a set of mechanistically-based assumptions regarding the xenobiotic interaction(s) in question (*1*). Recent reports that a wide variety of natural and synthetic compounds are capable of acting as steroid hormone ligand agonists and antagonists serve as timely examples of the need to advance mechanistically-based screening techniques to support human health and ecological risk assessments (*2*). In attempting to model hormone receptor−ligand interactions as one step in the mechanism of action for steroidal agonists and antagonists, the development of an efficient three-dimensional (3-D) approach to quantify chemical structure and similarity for large datasets is required.

Typical approaches to quantify 3-D similarity in the context of ligand−receptor interactions encompass pharmacophore (or toxicophore) search methods and receptor-site mapping (*3−8*). The selection of appropriate conformations and the proper alignment of structures to the lead compound are significant challenges with these methods. There are a number of good techniques for superimposing molecules (*9−12*). Typically, hundreds of alignments are explored to reach an optimum outcome, which if not carefully evaluated and explained in the context of a presumed mechanism of interaction with the receptor may result in models that violate the criteria of Topliss and Edwards (*13*) for causality in structure−activity models. Further, the use of the lowest energy conformers for flexible structures to assess similarity in pharmacophore search and receptor-mapping algorithms may be inappropriate because, in complex systems such as biological tissues and fluids, it is likely that the minimum energy conformer does not interact with the target. In fact the lowest energy, gas-phase conformations might be the least likely to interact with the solvent or macromolecules (*14*), and solvation and binding interactions could compensate for energy differences among the conformers of a chemical (*15−18*). In an attempt to address the issue of conformational flexibility, Prendergast et al. (*19*) reported an approach to identify specific conformers of ligands acting as antagonists to the angiotensin-II receptor, based on interatomic distances. All geometrically reasonable conformers were assessed; however, conformational energies were not evaluated, and energy minimization was not performed during the search.

The objective of the present study was to generalize the use of multiple conformers in an active analogue approach. The COmmon REactivity PAttern (COREPA) approach circumvents the problems of conformer alignment and selection, and initial assumptions concerning specific atoms/fragments in a pharmacophore are not an obligatory step. In addition, the method is not restricted to using interatomic distances, but implicitly defines the common reactivity pattern across any global and local reactivity parameter hypothesized to be associated with the specific biological endpoint under study. To illustrate the algorithm, the stereoelectronic requirements associated with the binding of a diverse set of ligands to the androgen receptor (AR) were defined. This particular analysis was undertaken, in part, because of recent concern that adverse effects on the health of humans and wildlife have been or currently are being elicited by chemicals that exert toxicity through agonistic or antagonistic interactions with steroid receptors (*2, 20*). In turn, this concern has lead to the call for developing techniques to support first-tier, or screening-level, hazard identifications for new and existing chemicals (*2*).

* To whom correspondence should be addressed. Telephone: 218-720-5500; fax: 218-728-5703; e-mail: bradbury.steven@epamail.epa.gov.

Materials and Methods

The methodology to elucidate chemical similarity 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. Elucidation of this common reactivity pattern within a set of toxicologically similar chemicals requires examination of the conformational flexibility of the compounds to evaluate molecular similarity in the context of the associated variability in specific stereoelectronic parameters.

(A) The COREPA Approach. The Algorithm. The principal steps of the algorithm can be summarized as follows:

Step 1. Definition of the Training Set of Chemicals. A defined subset of chemicals in the reaction series under investigation are selected as the training set. The training set can include either the most or least active chemicals, as defined by a user-imposed threshold of biological activity. This initial step establishes the extent of biological similarity among the chemicals from which stereoelectronic similarity will be discerned in the subsequent steps of the algorithm.

Step 2. Evaluation of Stereoelectronic Parameters Hypothesized To Be Associated with Biologically Similar Compounds. A restricted set of parameters, hypothesized to be associated with biological activity, are evaluated based on the normalized sum of dynamic similarity indices (see Dynamic Similarity Method) between each pair of molecules in the training set. For each parameter evaluated, the normalization is based on the sum of the pair-wise similarity index values divided by the number of pairs examined. The stereoelectronic parameters that provide the maximal measure of similarity among the chemicals in the training set are assumed to be most closely associated with the activity under consideration and are used in the subsequent step of the algorithm. To minimize the potential for creating chance similarity relationships, or similarity relationships that are unique to the training set, the algorithm is not employed by evaluating all possible stereoelectronic parameters, but only those that have a mechanistically-reasonable basis.

Step 3. Recognition of the Common Reactivity Pattern. For each stereoelectronic parameter identified in step 2, the conformer distributions of the chemicals from the training set are superimposed, and the parameter ranges common for conformers from all of the chemicals are identified. The distribution intersections (i.e., commonly populated parameter ranges) can be either discrete or continuous. Note, it is possible that for some of the parameters identified in step 2 there may not be a common range that incorporates at least one conformer from each chemical. The collection of common stereoelectronic parameter ranges defines the common reactivity pattern. In the context of local stereoelectronic ranges, it is essential to subsequently elucidate the associated specific atom types, fragments, and interatomic distances.

Conformer Generation. A primary aspect of the approach is to evaluate the conformational space for the chemicals under study using a number of conformers that can reasonably be assumed to represent the diversity of relevant stereoelectronic character for the biological process of interest. Conformer generation was based on a combinatorial procedure (*21*) that initiates from molecular topology and generates all conformers consistent with steric constraints (e.g., distances between non-bonded atoms, ring-closure limits, torsion resolution) and expert rules (likelihood of intramolecular hydrogen bonds, cis/trans, or ±isomers). This technique has been used in modeling ligand binding affinity to the estrogen (ER) and aryl hydrocarbon receptors (AhR) (*16, 17*). Of special note is that the technique incorporates the conformational flexibility of saturated cyclic molecular fragments, as opposed to other techniques that only explore conformational space formed by rotations around acyclic single bonds (*21*). In cases with strained cyclic structures,

the exhaustive generation of 3-D isomers usually requires less restrictive geometric constraints. As a result, some of the generated conformers could be distorted with respect to geometric parameters. In such cases an original strain minimization technique (pseudo-molecular mechanics, PMM) is applied based on a simple energy-like function where only the electrostatic terms are omitted (*21*). Geometry optimization can then be completed by more accurate force field or quantum-chemical methods.

Dynamic Similarity Method. The implementation of step 2 in the COREPA algorithm is based on a procedure described by Mekenyan et al. (*22*), where specific descriptors are evaluated across conformer data sets to elucidate those aspects of stereoelectronic structure most similar among compounds with similar biological activity (as defined in step 1 of the approach). For stereoelectronic parameters selected based on a known or hypothesized involvement in a biological process, parameter values are derived for all of the conformers for all of the chemicals within a data set. The resulting parameter range is then partitioned, and a frequency distribution is created for each compound that is based on the number of conformers that populate the parameter partitions. In cases where the parameter is an atomic or local stereoelectronic descriptor, such as charge, superdelocalizability, polarizability, etc., the frequency distributions are based on the number of atoms from all the conformers that populate the partitions. However, the correspondence of an atom to a specific conformer is maintained. Interatomic distances in molecules are also considered to be local parameters, and these frequency distributions are based on the number of distances from all the conformers located within specified distance partitions.

Following the establishment of conformer frequency distributions for each compound, a pattern-based similarity search is then employed, where frequencies of conformers derived from specific compounds are compared across the entire chemical and parameter data set. When superimposing chemical-specific conformer distributions for two compounds, the partitions can be occupied by conformers of one of the molecules or by the conformers of both molecules. In the current study, the difference in frequency distributions for two molecules was assessed by the Tanimoto coefficient S(AB), the related Hodgkin-Richards similarity metric, SH(AB), cosine, COS(AB), the Euclidean metric, DS(AB), and the Shannon information content, I(AB) (*23−29*). The index values for S(AB), SH(AB), and COS(AB) corresponding to maximum similarity and minimum similarity are 1 and 0, respectively. For DS(AB), maximum and minimum similarity values are 0 and infinity, respectively. I(AB) assesses conformers populating common partitions only and tends toward 0 when common partitions are evenly populated. The Supporting Information provides a more detailed description of these indices.

When comparing molecules, it is essential that the geometric constraints used to generate conformers be held constant across all chemicals in the data set. The use of different geometric constraints for chemicals in the data set could lead to a bias in the conformer distributions that results in similarity conclusions that are actually based on the conformer generation algorithms, rather than presumed differences in 3-D character.

(B) COREPA Analyses for AR Ligand Binding Affinity. AR Ligands and Binding Affinity. The AR ligands examined in this study are depicted in Figure 1 and consist of nine steroids, two synthetic steroids, and 17 non-steroids. AR binding affinities were obtained from Kelce et al. (*30*) and Waller et al. (*31*) and are based on a competitive binding assay using [^{3}H]RI881 [a radiolabled synthetic androgen; see Kelce et al. (*30*)]. This data set was selected because it represents a compilation of binding data generated within the same laboratory using consistent conditions, thereby
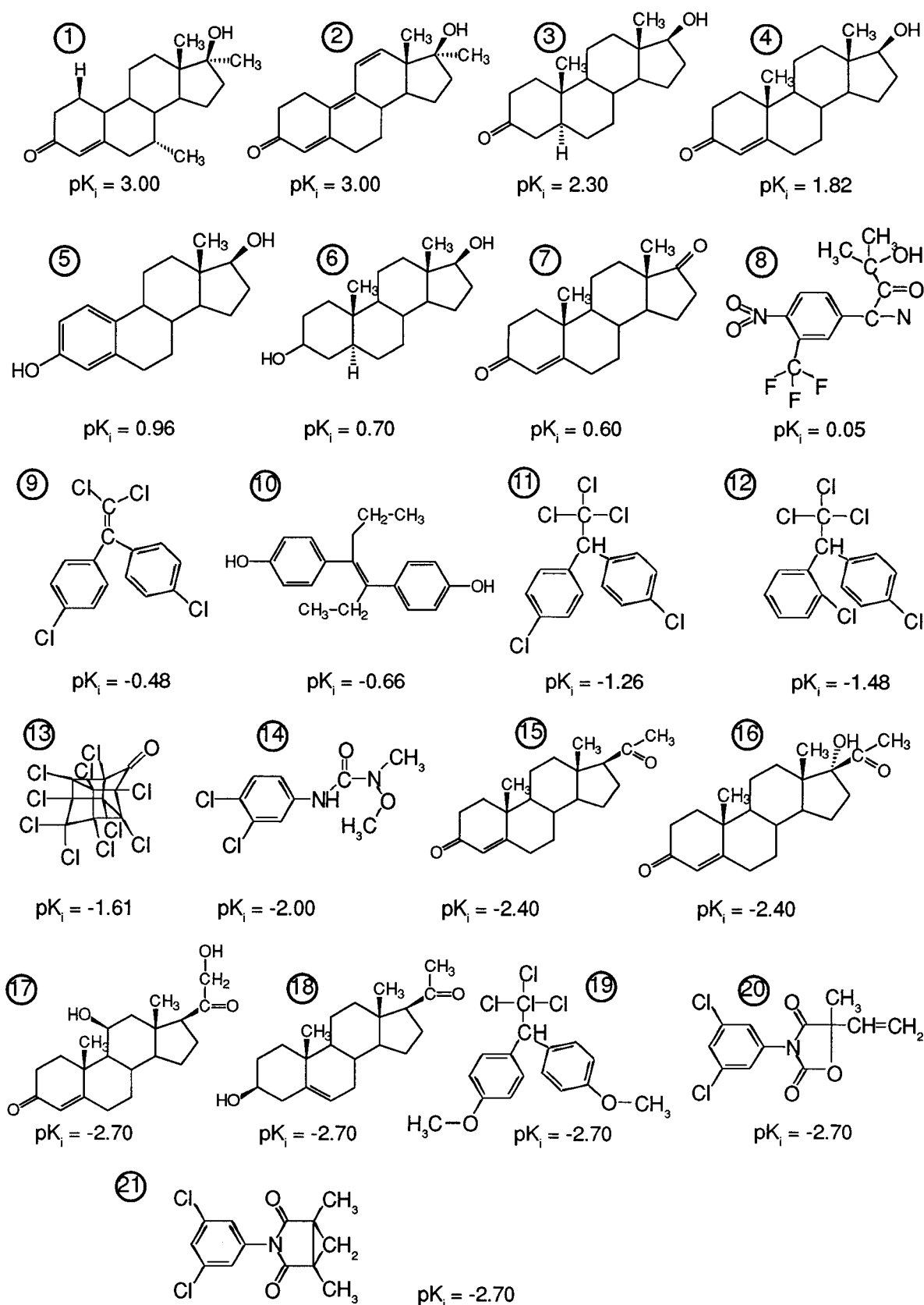
FIGURE 1. Structures of androgen receptor ligands used to establish training and validation sets (see Table 1 for compound names).

minimizing the contribution of biological variability to uncertainty in the modeling results. In addition, the data set represents several chemical classes, which include natural ligands as well as those representative of xenobiotics proposed to disrupt endocrine function (*30, 31*). The pK_i values (the negative log of the equilibrium dissociation constants;

provided by C. L. Waller) are listed in Table 1. In the following analyses, 2-{[(3,5-dichlorophenyl)carbamoyl]oxy}-2-methyl-3-butenoic acid (M1), 3′,5′-dichloro-2-hydroxy-2-methylbut-3-en anilide (M2), 3,5-dichlorobenzanilide 2-cyclopropane-carboxylic acid (P1), 2,2-bis(*p*-hydroxyphenyl)-1,1,1-tri-chloroethane (HPTE), hydroxylinuron, PCB153, and the

TABLE 1. Androgen Receptor Ligands, Observed Binding Affinities, Generated Conformers, and Ranges of RMS and ΔH_f° (Based on N_2 Conformers)

|		pK (μM)[a]	no. of conformers		parameter ranges	
compd no.	androgen receptor ligands	obsd	N_1	N_2	RMS	ΔH_f° [kcal/mol]
1	mibolerone	3.00	43	9	0.772−2.002	−117.866 to −106.652
2	methylrienolone	3.00	169	12	0.260−1.717	−68.008 to −52.750
3	5α-dihydrotestosterone	2.30	23	5	0.236−0.478	−143.529 to −134.074
4	testosterone	1.82	44	6	0.409−1.624	−115.286 to −104.969
5	estradiol	0.96	19	4	0.334−0.863	−107.616 to −96.898
6	5α-androstane-3α,17β-diol	0.70	17	4	0.131−0.541	−160.722 to −150.943
7	Δ[1]-androstenedione	0.60	15	5	0.587−1.816	−95.834 to −81.456
8	hydroxyflutamide	0.05	28	19	2.007−4.972	−213.829 to −195.491
9	*p,p*-DDE	−0.48	15	4	0.874−3.654	39.987 to 43.957
10	diethylstilbestrol (DES)	−0.66	126	21	1.909−7.950	−47.519 to −45.062
11	*p,p*-DDT	−1.26	19	5	1.312−6.785	18.832 to 20.424
12	*o,p*′-DDT	−1.48	24	14	1.917−7.351	21.096 to 32.743
13	kepone	−1.61	1	1	0	18.354
14	linuron	−2.00	27	26	1.483−4.911	−13.921 to −2.541
15	progesterone	−2.40	291	22	0.385−1.911	−105.648 to −88.658
16	17α-hydroxyprogesterone	−2.40	268	24	0.800−2.631	−142.232 to −126.488
17	corticosterone	−2.70	129	35	0.531−2.327	−191.442 to −172.466
18	pregnolone	−2.70	198	13	0.595−1.557	−127.281 to −119.877
19	methoxychlor	−2.70	31	20	2.264−6.126	−43.447 to −39.594
20	vinclozolin	−2.70	24	11	0.667−3.189	−57.180 to −47.413
21	procymidone	−2.70	16	6	1.590−4.995	−16.206 to −9.140

[a] Data obtained from ref 31.
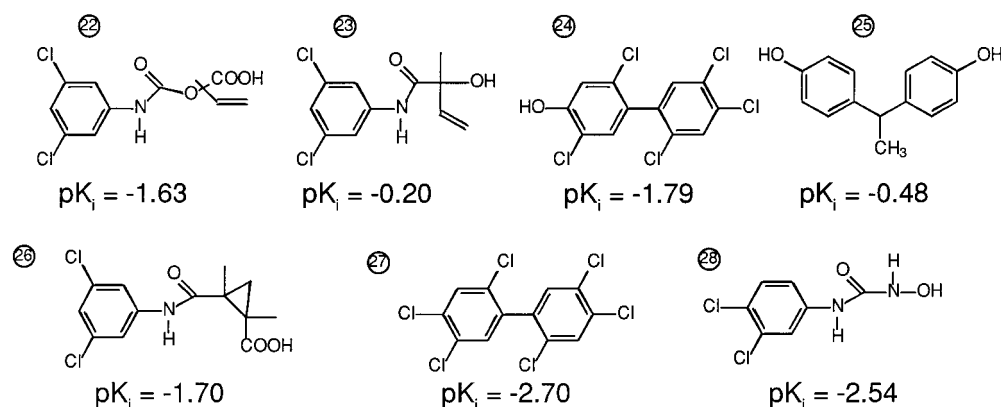
FIGURE 2. Structures of androgen receptor ligands with experimental binding affinities (*31*) used as an external evaluation set (22, 2-{[(3,5-dichlorophenyl)carbamoyl]oxy}-2-methyl-3-butenoic acid; 23, 3′,5′-dichloro-2-hydroxy-2-methylbut-3-en anilide; 24, the hydoxylated analog of PCB153; 25, 2,2-bis(*p*-hydroxyphenyl)-1,1,1-trichloroethane; 26, 3,5-dichlorobenzanilide 2-cyclopropanecarboxylic acid; 27, PCB153; 28, hydroxylinuron).

hydoxylated analog of PCB153 (Figure 2) were used as an external validation data set to evaluate similarity models derived in the study.

AR Ligand Conformations. In generating conformers for the AR ligands, specific geometric constraints were imposed for different types of bonds in the chemical data set. The torsion resolution (TR) around 'saturated' (SP3−SP3) bonds was chosen to be 120° using an initial torsion angle of 60° with respect to the plane of the preceding three atoms. For aromatic fragments and incident bonds, the TR was set at 180°, with an initial torsion angle of 0°, to minimize the combinatorics. For all non-steroid chemicals, 1.5 Å was set as the distance between non-bonded atoms, while 1.2−1.8 Å was the range imposed for ring closure. The same geometric constraints were imposed for the acyclic moieties of the steroids. Due to the rigidity of the natural steroids and their derivations, less restrictive geometric constraints for ring closures (1.0−2.5 Å and torsion resolution of 60°) were imposed for their cyclic moieties to generate a sufficiently large set of conformations with the same stereospecificity as the natural enantiomers with B/C trans and C/D trans ring fusion. Each of the generated conformers were submitted to PMM optimization. Subsequently, the conformational de-

generacy of the isomers was detected (within a 30° range of torision angle difference) due to molecular symmetry and geometry convergence. The number of conformations generated under these constraints is listed in Table 1 (denoted as N_1). Subsequent geometry optimization of the conformers was employed with MOPAC 7 (*32*), using the AM1 Hamiltonian with the key words 'PRECISE' and 'NOMM'. As a result of the optimizations, some of the conformations quenched into the same energy minima, which reduced the number of conformers (denoted as N_2 in Table 1). Finally, these conformers were screened to eliminate any structures with ΔH_f° values that were 20 kcal/mol or higher than that calculated for the conformer associated with the 'absolute' energy minimum (see Results and Discussion for an explanation of this threshold). In the current study, none of the conformers exceeded this threshold.

As an illustration, after energy optimization (first by the PMM technique and then by the quantum chemical approach), the 19 conformations of estradiol (E$_2$) originally generated (N_1 in Table 1) quenched into four conformers having differences of at least 30° for at least one pair of their torsion angles (based on the 30° threshold for conformational degeneracy). The range in ΔH_f° values for the energy minima

I -107.616 kcal/mol

II -106.053 kcal/mol

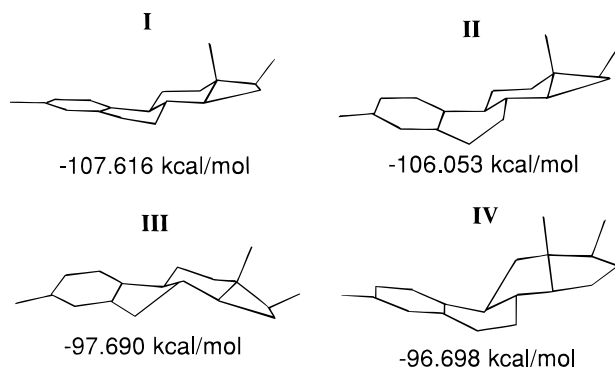**III** -97.690 kcal/mol

IV -96.698 kcal/mol

FIGURE 3. Generated E_2 conformers having the same stereochemistry as the natural enantiomer with its B/C trans and C/D trans fusions. Conformational degeneracy of the optimized conformations was detected at a 30° threshold for torsion angle differences. The conformers are energy minima obtained by making use of force field and quantum chemical approaches. The AM 1 calculated formation enthalpies are presented. Conformers I and II are those predicted from the simulated annealing search method by Wiese and Brooks (*18*).

associated with these four E_2 conformers (Figure 3) was -107.616 to -96.898/kcal/mol ($\Delta\Delta H_f° = 10.718$ kcal/mol), which is comparable with the free energy of E_2 binding to the ER (e.g., ΔG of -12.1 kcal/mol at 4 °C; $K_a = 3.7 \times 10^9$ M^{-1}; *18*). As can be seen in Figure 3, energy differences between some of the structures are relatively small, e.g., 1.56 kcal/mol between conformers I and II and 0.593 kcal/mol for III and IV. The lowest energy E_2 conformer corresponds to the previously reported X-ray structure (*33, 34*). The second conformer, with formation enthalpy of -106.053 kcal/mol, has also been predicted from a simulated annealing search method described by Wiese and Brooks (*18*). The differences between conformers I and II reside entirely in altered B-ring conformations, which is a distorted $7\alpha,8\beta$ half-chair for E_2-I and the boat conformation for E_2-II. The third conformation predicted in the present work is similar to E_2-II, however, a distorted boat is generated for the B-ring, which produces a twisting of the C- and D-rings. This distortion results in a 8.36 kcal/mol increase in $\Delta H_f°$, relative to the conformationally similar E_2-II. At higher thresholds for degeneracy (e.g., 60°), conformer III is identical to E_2-II. The B- and C-ring conformations in E_2-IV (B-ring as a distorted boat and C-ring as a boat conformation) have entirely different configurations compared to E_2-III, although they are associated with small differences in $\Delta H_f°$ values.

Table 1 lists the range of root mean square (RMS) differences for each compound, based on comparisons of each conformer with the lowest energy structure. Smaller RMS ranges were associated with steroid derivatives, consistent with their greater rigidity. For example, RMS ranges of 0.722–2.002, 0.334–0.863, and 2.007–7.972 were derived for mibolerone (**1**), estradiol (**5**), and hydroxyflutamide (**8**), respectively.

Molecular Descriptors. The global and local electronic descriptor pool used in the present study was restricted to parameters hypothesized to be associated with AR binding affinity based on previous analyses by Waller et al. (*31*), as well as other studies using a variety of model receptors (*16, 17, 35, 36*). Stereoelectronic parameters were calculated with MOPAC 7 (*32*), augmented by a computing module that provides additional reactivity descriptors, using the AM1 all-valence electron, semi-empirical Hamiltonian. The electronegativity (EN), dipole moment (μ), volume polarizability (VolP) energy of frontier orbitals (E_{HOMO} and E_{LUMO}), and the electronic gap ($E_{HOMO-LUMO}$) were used as global electronic descriptors, whereas the atomic charges (q_i), frontier atomic charges (f_i^{HOMO} and f_i^{LUMO}), and donor and acceptor super-

delocalizability indices (S_i^E and S_i^N; *37–39*), as well as atomic self-polarizabilities (p_{ii}) were calculated as local electronic indices (*i* denotes a specific atom in a molecule). In the present study, when searching common patterns based upon local parameter distributions, the atomic reactivity indices were not restricted to specific rings in steroidal or non-steroidal derivatives.

Conformers were also screened based on the steric descriptors GW (sum of geometric distances; *40*), L_{max} (the greatest interatomic distance), d_{ij} (steric distance between atoms *i* and *j*), and planarity (the normalized sum of torsion angles in a molecule; *16*). These descriptors were selected because hydrophobicity, steric bulk, and size constraints have been reported as important criteria in predicting and interpreting ligand binding for nuclear steroid receptors (*17, 31, 35, 41–43*). Finally, volume polarizability (VolP), defined as a sum of atomic self-polarizabilities and thus the averaged ability of a compound to change electron density at its atoms during chemical interactions (*38, 44*), was selected as a physicochemical descriptor. Lower values of VolP (VolP > 0) reflect higher charge localizations and more polarizable (less lipophilic) molecules (*44*). This descriptor was used based on previous observations, suggesting that more polarizable polychlorinated hydroxybiphenyls ligand conformers generally had greater binding affinities to ER (*17*).

COREPA Analyses. To initially illustrate the approach, the eight conformers with greatest binding affinity and the eight conformers with the lowest binding affinity were selected as training sets (step 1) and carried through steps 2 and 3, with 20 partitions used in step 3. Next, the influence of biological similarity on quantifying chemical similarity was illustrated in step 3. Analyses, restricted to local atomic charges and local distances, were used in training sets containing four, six, or eight of the most active compounds (pK_i value ranges of 3.00–1.82, 3.00–0.70, and 3.00–0.05, respectively) with 10, 20, 30, or 40 partitions. The common reactivity pattern obtained with the six most active isomers was subsequently evaluated based on the ability to differentiate subsets of the remaining compounds in Table 1 as well those in the external evaluation set of structures (Figure 2). To facilitate a summary of the results, hereafter the terms high and low activity are used to denote high and low ligand binding affinity. However, the term activity should not be construed to impart any conclusion as to biocharacter or potency of the ligands.

Results and Discussions

Conformational Flexibility and Electronic Structure. The range of $\Delta H_f°$ values for conformers of any chemical was less than 20 kcal/mol (see column N_2 in Table 1). Assuming a free energy of binding to steroid hormones in the range of -10 to -15 kcal/mol [e.g., -12.1 kcal/mol for the binding of E_2 to the ER (*18*)], the 20 kcal/mol threshold for $\Delta\Delta H_f°$ was assumed to result in an energetically reasonable set of conformations, especially given the extent to which energy provided during ligand binding could facilitate conformational transformations (*17, 18*). To provide a sense of variability in 3-D shape associated with the conformers used in the analysis, relevant torsional angles and non-bonded interatomic distances associated with the lowest and highest energy structures are presented in the Supporting Information.

For a given compound, conformers within the specified range of $\Delta\Delta H_f°$ often exhibited significant variation in potentially relevant electronic descriptors, as summarized within Table 1 of the Supporting Information. For example, conformers of hydroxyflutamide (**8**) had a range 0.63 eV for E_{LUMO}, 1.09 eV for E_{HOMO}, 0.61 eV for $E_{HOMO-LUMO}$, 0.011 (au)²/eV for VolP, and 6.51D for μ. Similar variations were observed for the non-steroidal compounds **9**, **10**, **14**, **20**, and **21**. The parameter ranges for the steroids, while smaller (likely due

TABLE 2. Average Tanimoto (S_{av}), Euclidean Distance (DS_{av}), Shannon Information Content (I_{av}), Hodgkins−Richards Similarity (SH_{av}), and Cosine (COS_{av}) Metrics Normalizing Pairwise Similarity between Eight Most Active (pK_i Values > 0.0) and Eight Least Active (pK_i Values $\leq$ −2.00) Androgen Ligands (See Table 1) in Context of Stereoelectronic Descriptors Hypothesized To Be Related to Ligand Binding to AR

compd no.	parameters	eight most active ligands					eight least active ligands				
		S_{av}	DS_{av}	I_{av}	SH_{av}	COS_{av}	S_{av}	DS_{av}	I_{av}	SH_{av}	COS_{av}
1	ΔH_f°	0.000	0.354	0.000	0.000	0.000	0.006	0.376	0.063	0.000	0.000
2	GW	0.000	0.354	0.000	0.000	0.000	0.000	0.365	0.000	0.000	0.000
3	L_{max}	0.000	0.354	0.000	0.000	0.000	0.016	0.477	0.162	0.000	0.000
4	planarity	0.000	0.354	0.000	0.000	0.000	0.033	0.463	0.252	0.000	0.000
5	VoIP	0.000	0.354	0.000	0.000	0.000	0.000	0.354	0.000	0.000	0.000
6	$E_{HOMO-LUMO}$	0.000	0.456	0.000	0.000	0.000	0.023	0.686	0.167	0.000	0.000
7	EN	0.000	0.399	0.000	0.000	0.000	0.034	0.600	0.438	0.314	0.883
8	E_{LUMO}	0.001	0.365	0.018	0.000	0.000	0.017	0.654	0.198	0.000	0.000
9	μ	0.018	0.779	0.109	0.144	0.447	0.112	0.433	0.862	0.761	0.991
10	E_{HOMO}	0.024	0.934	0.226	0.545	0.897	0.029	1.191	0.310	0.067	0.125
11	S_i^N	0.068	0.357	1.689	0.702	0.859	0.109	0.366	1.951	0.866	0.997
12	S_i^E	0.082	0.369	1.719	0.545	0.984	0.127	0.363	1.989	0.793	0.973
13	f_i^{HOMO}	0.083	0.377	1.721	0.358	1.000	0.149	0.362	2.023	0.846	0.998
14	f_i^{LUMO}	0.085	0.401	1.710	0.640	1.000	0.153	0.361	2.031	0.986	0.994
15	q_i	0.089	0.395	1.736	0.493	0.925	0.132	0.382	1.995	0.693	1.000
16	π_{ii}	0.089	0.365	1.737	0.460	0.954	0.120	0.365	1.978	0.873	0.941
17	d_{ii}	0.092	0.374	2.523	0.437	0.969	0.120	0.408	2.708	0.327	0.671

to their rigidity) are also noteworthy. For example, conformers of methyltriendone (**2**) had a range of 0.296 eV for E_{LUMO}, 0.175 eV for E_{HOMO}, 0.463 eV for $E_{HOMO-LUMO}$, 0.01 (av)^2/eV for VoIP, and 1.60D for μ. The observation that relatively small energy differences between conformers can be associated with significant variations in electronic structure highlights the necessity of including all energetically-reasonable conformers when defining common reactivity patterns.

Illustration of the COREPA Algorithm. To illustrate the COREPA algorithm, training sets were established (step 1) by selecting the eight most active AR ligands (pK_i values > 0.0) and comparing their reactivity patterns to the eight least active ligands (pK_i values $\leq$ −2.0). Pairwise similarity was assessed between chemicals within each training set by employing the 3-D dynamic similarity method (step 2). Potentially relevant reactivity parameters were identified for the most and least active compounds based on S_{av}, SH_{av}, COS_{av}, DS_{av}, and I_{av} values, normalized over all pairs of compounds within a training set (see Table 2; steric and electronic parameters are listed based on values of S_{av}).

For all indices, local steric and electronic parameter distributions typically exhibited greater degrees of similarity within each training set than did global parameters. Trends in I_{av} were found to parallel S_{av} (r^2 = 0.96 and 0.88 for the most and least active ligands, respectively), as did trends between SH_{av} and COS_{av} (r^2 = 0.98 and 0.77). Values for DS_{av} showed some degree of orthogonality to the other indices. However, in general, higher values of S_{av}, SH_{av}, COS_{av}, and I_{av} corresponded to lower values of DS_{av}. As defined in eqs 4 and 5 in the Supporting Information, DS_{av} may overestimate differences in flexibility between molecules, i.e., the differences in population intensity with each partition. Thus, even in cases where conformers of both molecules occupy common and broad parameter partitions, the DS_{av} value may still be large (suggesting high dissimilarity), possibly due to large differences in population intensities within the partitions.

The data summarized in Table 2 suggest that the eight most active and inactive AR ligands have the greatest similarity to members within each respective training subset in terms of distributions of charges (q_i), steric distances (d_{ij}), donor delocalizabilities (S_i^E), frontier charges on HOMO (f_i^{HOMO}) and atom polarizabilities (π_i). To establish common reactivity patterns (step 3), conformer frequency distributions of compounds from each of the two training sets were subsequently examined across all local stereoelectronic descriptors presented in Table 2, with results based on q_i, presented in
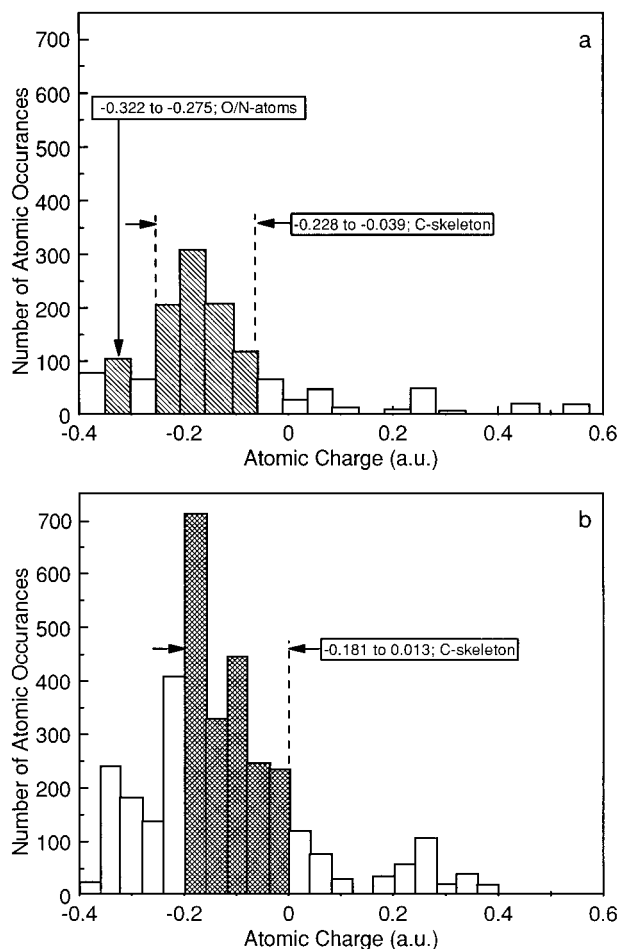

FIGURE 4. Common parameter ranges and conformer frequency distributions based on atomic charges (q_i) with 20 parameter partitions for (a) the eight most active androgen receptor ligands (pK_i values > 0.0; diagonal cross-hatch); (b) the eight least active androgen receptor ligands (pK_i values $\leq$ −2.00; full cross-hatch). See Table 1 for compound names and associated binding affinities.

Figure 4. In these analyses, the number of parameter partitions was set at 20 for each descriptor.

Even though conformer distributions from the most active and least active sets of compounds overlap, the subset of
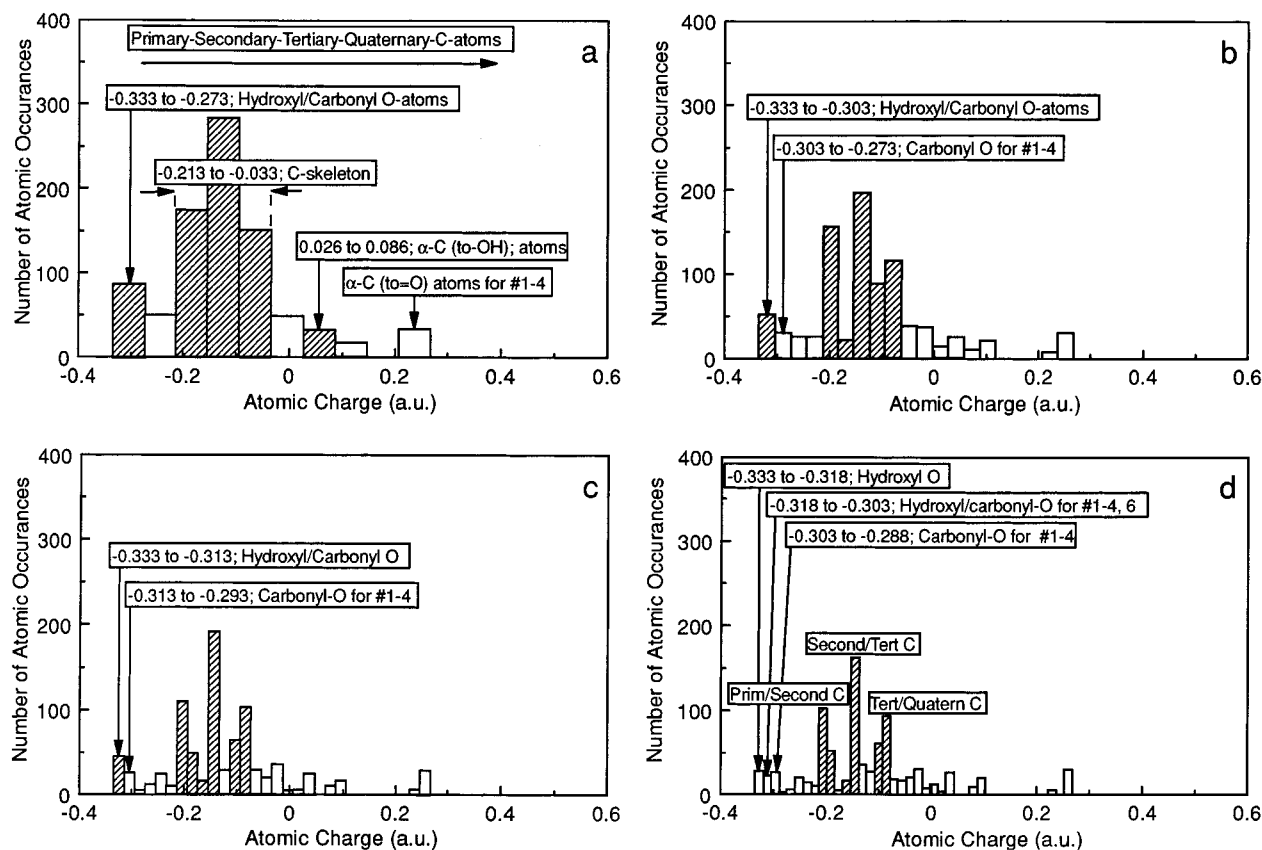
FIGURE 5. Charge frequency (q_i) distribution of the six most active androgen receptor ligands (pK_i values $\geq$0.70; see Table 1) with 10 (a), 20 (b), 30 (c), and 40 (d) partitions. Diagonal cross-hatch denotes ranges populated by at least one conformer from each ligand.

common partitions that contains conformers from each of the active compounds does not overlap with partitions containing conformers from each of the least active compounds. Thus, the common atomic charge pattern associated with the eight most active ligands (Figure 4a) deviates significantly from that of the eight least active chemicals (Figure 4b). For the most active compounds, common partitions (the diagonal cross-hatched bars in Figure 4) corresponded to oxygen and nitrogen (8) atoms (−0.322 to 0.275 au) and carbon atoms (−0.228 to −0.039 au) of the steroid template (denoted as 'C-skeleton'). While a common range of charges was observed for C-skeleton carbons in the least active compounds (full cross-hatched bars in Figure 4), no common range corresponding to oxygens was identified (Figure 4b). Similar observations were made for reactivity patterns based on atom polarizabilities where common ranges due to oxygen atoms (0.030−0.032 au^2/eV), carbon atoms next to oxygens (α-C atoms; 0.044−0.046 au^2/eV), and C-skeleton atoms (0.048−0.050 au^2/eV) were observed for the most active compounds (data not shown). While a similar range of oxygen polarizability was noted for the least active compounds, a common range for α-C atoms was not observed, and the range for C-skeleton atoms was significantly wider. A comparison of reactivity patterns based on interatomic distances also showed significant differences between active and inactive compounds. The common partition that included the largest interatomic distances for the most active compounds was 9.75−10.20 Å (data not shown). It should be noted, however, that a range of 10.64−11.10 Å, formed by the largest interatomic distances for conformers of the seven most active compounds, included oxygen−oxygen interatomic distances only. A common range based on distances between oxygens was not apparent for the least active compounds. For this training set, a distance range of 8.72− 9.26 Å, corresponding to the largest interatomic distances between carbon−carbon and carbon−oxygen atoms, was observed.

Previous modeling (17, 31, 35, 41) and structural studies (35, 42, 43) of nuclear steroid receptors have indicated the importance of steric constraints and size in predicting and interpreting ligand binding. For example, Waller et al. (31) reported that, for steroid derivatives, increases in steric bulk (size) in the vicinity of the B-ring off of the C6 and C7 atoms, the C-ring off of the C11 and C12 atoms, and the D-ring off the C17 atom were associated with increased binding affinity. Consistent with these observations, 3-D models of steroid receptor ligand binding domains suggest specific spatial arrangements of amino acid residues thought to be associated with A- and D-ring ligand interactions (35, 42, 43). Thus, the finding in this study that interatomic distances are similar for active ligands likely is associated with size constraints in the binding domain and/or indirectly associated with local steric characteristics.

The observation that active AR ligands have specific local electronic descriptor parameter ranges associated with greater negative charge and electron donating character also is consistent with previous research for steroid receptors. Studies by Waller et al. (31) suggested that increased negative charge in the vicinity of the C3 atoms of the A-ring and C17 substituents of the D-ring are associated with increased ligand binding affinity of steroidal compounds. These findings also are consistent with those of Waller et al. (41), Bradbury et al. (17), and VanderKuur et al. (36) that indicated similar characteristics in ER ligands. While recent structural studies of the ligand binding domains of nuclear receptors suggest that these sites are largely hydrophobic (35, 42, 43), there is also evidence of specific polar residues that may be associated with hydrogen bonding and weak polar interactions (35, 43). In turn, different residues within the binding pockets of different nuclear receptors are presumed to be responsible for determining ligand specificity (43).

Influence of Training Set Size and Number of Partitions on Algorithm Behavior. Reactivity patterns based on atomic charge and interatomic distances for combinations of training
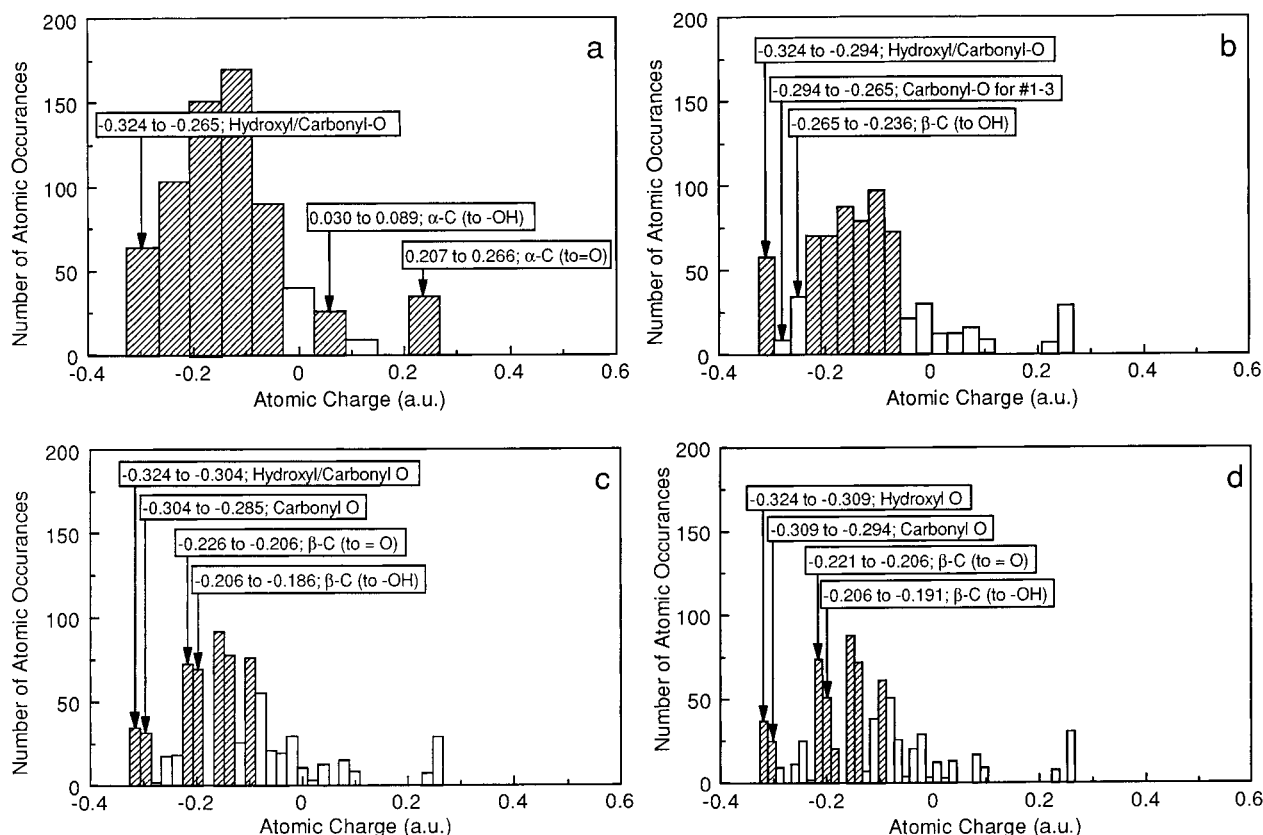
FIGURE 6. Charge frequency (q_i) distribution of the four most active androgens (pK_i values ≥ 1.82; see Table 1) with 10 (a), 20 (b), 30 (c), and 40 (d) partitions. Diagonal cross-hatch denotes ranges populated by at least one conformer from each ligand.

set size (i.e., ranges of pK_i of values) and number of partitions used to define frequency distributions were established to illustrate how variations in biological activity (i.e., different assumptions of biological similarity in terms of pK_i values) could be related to measures of chemical similarity.

When using a training set consisting of the eight most active ligands (pK_i values of 3.00−0.50), a common partition based on oxygen charges was not identified, nor was there a consistency in atom types associated with interatomic distances (data not shown). To illustrate how definitions of biological similarity influence interpretations of chemical similarity, training sets with six (pK_i values of 3.00−0.96) or four (pK_i values of 3.00−1.82) ligands were also evaluated. With training sets of six and four ligands, 'oxygen' windows appeared in all reactivity patterns independent of the number of partitions. Moreover, with this set of ligands and range of binding affinity, the specificity of reactivity patterns based on atomic charge increased with a decrease in the size of the training set (i.e., with an increase in binding affinity), as noted in Figures 5 and 6. In the analysis based on interatomic distance, a consistency in distribution characteristics associated with training set size and partition number was also observed. With six ligands in the training set, the largest common distances ranged from 10.22 to 11.10 Å, from 10.66 to 11.10 Å, from 10.80 to 11.10 Å, and from 9.78 to 10.00 Å for 10, 20, 30, and 40 partitions, respectively. With four ligands, the distance ranges were 10.11−10.97 Å, 10.54−10.97 Å, 10.68−10.97 Å, and 10.76−10.97 Å for 10, 20, 30, and 40 partitions, respectively. When using partitions of 20 or more, the largest interatomic distance ranges included oxygen−oxygen distances only. It must be stressed that these results are specific to the training sets employed. Different definitions of biological activity will result in different common reactivity patterns.

Evaluation of Reactivity Patterns. An initial evaluation of the ability of the algorithm to screen 'unknown' compounds for AR binding affinity was obtained through an analysis of

9−21 in Table 1, none of which were used in evaluating reactivity patterns of active compounds. These 13 compounds were divided into two subsets to better illustrate the characteristics of the screening approach. The first subset included **9−13** (pK_i values of −0.48 to −1.61) and was represented by 45 conformers. The second subset was comprised of **14−21** (pK_i values of −2.00 to −2.70) and was represented by 157 conformers.

A one-step screening was employed that simultaneously incorporated charge and distance components of the reactivity patterns. On the basis of the analyses described previously, stable reactivity patterns for atomic charge and interatomic distance could be discerned for the four- and six-membered training sets. Using the six-membered training set, reactivity patterns were defined to determine whether ligands with pK_i values less than 0.70 could be discriminated. An interatomic distance range of 10.2−11.1 Å was identified as a distance screen typically associated with oxygen−oxygen distances between the A- and D-rings (based on 10 partitions). Alternatively, a distance range of 10.7−11.1 Å, based on 20 partitions, could be viewed as a more restrictive range. The common range of −0.333 to −0.303 au, analyzed with 20 partitions, was associated with hydroxyl and carbonyl oxygens and assumed to be the less restrictive charge range. A range of −0.333 to −0.313 au analyzed with 30 partitions and associated with hydroxyl or carbonyl oxygens was incorporated as a more restrictive range.

With the simultaneous fulfilment of interatomic distance and atomic charge criteria, the objective was to determine whether or not the algorithm could discriminate ligands with pK_i values of less than −0.48 or −2.00 from those with pK_i values of 0.70 or greater. All distance and charge ranges were assessed against 'wildcard' atoms in the validation sets (i.e., non-hydrogen atom types were not specified in the compounds when performing the screens), even though these parameter ranges were associated with oxygen−oxygen distances and oxygen charges. Screens based on wildcard

atoms were undertaken to illustrate the ability of the algorithm to assess similarity without the need to pre-determine a pharmacophore or establish an alignment against a lead, or template, molecule. Thus, the interatomic distances and charges that were assessed in the screenings were not, for example, specified to be hydroxyl or carbonyl oxygen, nor was an automated or manual assignment of A- and/or D-ring steroidal counterparts required.

After applying the least restrictive distance/charge requirements (10.2−11.1 Å and −0.333 to −0.303 au) to the first set of compounds (**9**−**13**), no conformers were identified, consistent with the pK_i values for these ligands. Employing the more restrictive distance/charge requirement (10.7−11.1 Å and −0.333 to −0.313 au) to **14**−**21** resulted in the identification of one conformer for **17** (pK_i = −2.70), i.e., an incorrect prediction of a pK_i of at least 0.70. However, employing a reactivity pattern based on the four most active ligands (with 30 partitions, 10.7−11.0 Å and −0.324 to −0.304 au, based on hydroxyl oxygens) this compound was identified as having a pK_i less than 1.82.

Finally, the reactivity pattern-based screening approach was assessed against a validation set comprised of the seven compounds in Figure 2. The conformer generation routine and subsequent quantum chemical optimization produced 132 conformers for the seven compounds (all within 20 kcal/mol of the lowest energy geometries). All of these compounds were properly discriminated using both the most and least restrictive distance/charge requirements. Thus, all seven compounds were identified as ligands likely to exhibit a pK_i binding affinity less than 0.70.

In contemplating the application of this approach with large databases of 3-D structures, it is reasonable to assume that only a single conformer per compound would be available and that the generation of energetically-reasonable conformers for each compound in an entire dataset would be computationally impractical. In these situations, we envision the use of an initial, less restrictive screening strategy that assesses single conformers and that would be designed to minimize the percentage of false negatives (i.e., compounds incorrectly predicted to be below a specified threshold). A prescreen based on the use of the 'tweak' technique (*45*) could also be used to manipulate rotatable bonds in an attempt to generate potentially active conformers from a single starting conformation. A second series of more refined screens based on sets of energetically-reasonable conformers, which requires the more time-consuming conformational analyses, would then be implemented. This type of an analogue search strategy requires that the 'stability' of reactivity patterns be carefully considered when selecting criteria for identifying active compounds. The strengths and weaknesses of such search strategies are currently being investigated.

Acknowledgments

Supporting Information Available

Text, equations, and references describing the dynamic similarity method as well as two tables that provide ranges for stereoelectronic parameters for conformers of **1**−**21** and torsional angles and non-bonded interatomic distances for steroid derivatives (6 pp) will appear following these pages in the microfilm edition of this volume of the journal. Photocopies of the Supporting Information from this paper or microfiche (105 × 148 mm, 24× reduction, negatives) may be obtained from Microforms Office, American Chemical Society, 1155 16th St. NW, Washington, DC 20036. Full bibliographic citation (journal, title of article, names of authors, inclusive pagination, volume number, and issue number) and prepayment, check or money order for $16.50 for photocopy ($18.50 foreign) or $12.00 for microfiche ($13.00 foreign), are required. Canadian residents add 7% GST. Supporting Information is also available via the World Wide Web at URL http://www.chemcenter.org. Users should select Electronic Publications and then Environmental Science and Technology under Electronic Editions. Detailed instructions for using this service, along with a description of the file formats, are available at this site. To download the Supporting Information, enter the journal subscription number from your mailing label. For additional information on electronic access, send electronic mail to si-help@acs.org or phone (202)872-6333.

Literature Cited

(1) Bradbury, S. P. *SAR QSAR Environ. Res.* **1994**, *2*, 89−104.
(2) Ankley, G. T.; Johnson, R. D.; Detenbeck, N. E.; Bradbury, S. P.; Toth, G.; Folmar, L. C. *Rev. Toxicol.* **1997**, *1*, 231−267.
(3) Marshall, G. R. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; Escom: Leiden, 1993; pp 80−116.
(4) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(5) Goodford, P. J. *J. Med. Chem.* **1993**, *28*, 849−857.
(6) Buchheit, K.-H.; Gamse, R.; Giger, R.; Hoyer, D.; Klien, F.; Kloppner, E.; Pfannkuche, H.-J.; Mattes, H. *J. Med. Chem.* **1995**, *38*, 2326−2330.
(7) Hahn, M. *J. Med. Chem.* **1995**, *38*, 2080−2090.
(8) Hahn, M.; Rogers, D. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(9) Kearsely, S. K.; Smith, G. M. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.
(10) Blaney, J. M.; Dixon, J. S. *Perspect. Drug Design* **1993**, *1*, 301−319.
(11) Perkins, T. D.; Deam, P. M. *J. Comput-Aided Mol. Design* **1993**, *7*, 155−172.
(12) Martain, Y. C.; Bures, M. G.; Danahar, E. A.; DeLazzar, J.; Lico, I.; Pavlik, P. A. *J. Comput.-Aided Mol. Design* **1993**, *7*, 83−102.
(13) Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(14) Eliel, E. L. In *Chemical Structures, Vol. I*; Warr, W. A., Ed.; Springer: Berlin, Germany, 1993; pp 1−8.
(15) Mekenyan, O. G.; Ivanov, J. M.; Veith, G. D.; Bradbury, S. P. *Quant. Struct.-Act. Relat.* **1994**, *13*, 302−307.
(16) Mekenyan, O. G.; Veith, G. D.; Call, D. J.; Ankley, G. T. *Environ. Health Perspect.* **1996**, *104*, 1302−1309.
(17) Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. *Environ. Chem. Toxicol.* **1996**, *15*, 1945−1954.
(18) Wiese, T,; Brooks, S. C. *J. Steroid Biochem. Mol. Biol.* **1994**, *50*, 61−72.
(19) Prendergast, K.; Adams, K.; Greenlee, W. J.; Nachbar, R. B.; Patchett, A. A.; Underwood, D. J. *J. Comput.-Aided Mol. Design* **1994**, *8*, 491−512.
(20) Kavlock, R. J.; Daston, G. P.; DeRosa, C.; Fenner-Crisp, P. Gray, L. E.; Kaatari, S.; Lucier, G.; Luster, M.; Mac, M. J.; Maczka, C.; Miller, R.; Moore, J.; Rolland, R.; Scott, G; Sheehan, D. M.; Sinks, T.; Tilson, H. A. *Environ. Health Perspect.* **1996**, *104*, 715−740
(21) Ivanov, J. M.; Karabunarliev, S. H.; Mekenyan, O. G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 234−243.
(22) Mekenyan, O. G.; Ivanov, J. M.; Karabunarliev, S. H.; Hansen, B.; Ankley, G. T.; Bradbury, S. P. A new approach for estimating three-demensional similarity that incorporates molecular flexibility. In *Proceedings of the 7th International Workshop on QSAR in Environmental Sciences*; Chen, F., Ed.; SETAC: Pensacola, FL, 1997 (in press).
(23) Everitt, B. *Cluster Analysis*; Halstead-Heinemann: London, 1980.
(24) Jakes, S. E.; Willett, P. *J. Mol. Graphics* **1986**, *4*, 12−20.
(25) Pepperrell, C. A.; Willett, P. In *Chemical Structures. The International Language of Chemistry*; Ware, W. A., Ed.; Springer: Berlin, 1993; Vol. 2, pp 377−382.
(26) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kentz, I. D. *J. Comput.-Aided Mol. Design* **1995**, *9*, 1−12.
(27) Bures, M. G.; Martain, I. C.; Willett, P. In *Topics in Stereochemistry*; Eliel, E. L., Wilen, S. H., Eds.; Wiley: New York, 1994; Vol. 1, pp 467−511.
(28) Holliday, J. D.; Ranade, S. S.; Willett, P. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(29) Shannon, C.; Weaver, W. In *The Mathematical Theory of Communication*; University of Illinois: Urbana, 1949.

(30) Kelce, W. R.; Monosson, E.; Gamasik, M. P.; Laws, S. C.; Earl, G. L., Jr.; *Toxicol. Appl. Pharmacol.* **1994**, *126*, 276−285.

(31) Waller, C. L.; Juma, B. W.; Gray, E. L., Jr.; Kelce, W. R. *Toxicol. Appl. Pharmacol.* **1996**, *137*, 219−227.

(32) Stewart, J. J. P. *MOPAC: A general molecular orbital packages*; Version 7.0. Software; Quantum Chemistry Program Exchange 455; University of Indiana: Bloomington, IN, 1995.

(33) Busetta, B.; Courseille, C.; Hospital, G. M. *Acta Crystallogr. B* **1972**, *28*, 1349−1351.

(34) Bussetta, B.; Hospital, G. M. *Acta Crystallogr. B* **1972**, *28*, 560−567.

(35) Anstead, G. M.; Carlson, K. E.; Katzenellenbogen, J. A. *Steroids* **1997**, *62*, 268−303.

(36) VanderKuur, J. A.; Wiese, T.; Brooks, S. C. *Biochemistry* **1993**, *32*, 7002−7008.

(37) Schuurmann, G. *Quant. Struct.-Act. Relat.* **1990**, *59*, 326−333.

(38) Mekenyan, O. G.; Karabunarliev, S. H.; Ivanov, J. M.; Dimitrov, D. N. *Comput. Chem.* **1994**, *18*, 172−177.

(39) Mekenyan, O. G.; Veith, G. D. *SAR QSAR Environ. Res.* **1994**, *2*, 129−143.

(40) Mekenyan, O. G.; Peitchev, D.; Bonchev, D. G.; Trinajstić, N.; Bangov, I. *Arzneim. Forsch.* **1986**, *36*, 176−183.

(41) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, E. L., Jr. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(42) Wurtz, J. M.; Bourguet, W.; Renaud, J. P.; Vivat, V.; Chambon, P.; Moras, D.; Gronemeyer, H. *Nature Struct. Biol.* **1996**, *3*, 87−94.

(43) Goldstein, R. A.; Katzenellenbogen, J. A.; Luthey-Schulten, Z. A.; Seielstad, D. A.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9949−9953.

(44) Lewis, D. F. V. *J. Comput. Chem.* **1989**, *10*, 145−151.

(45) Hurst, T. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.