J-CAMD 291

Replacement of steric 6–12 potential-derived interaction energies by atom-based indicator variables in CoMFA leads to models of higher consistency

Romano T. Kroemer* and Peter Hecht

Sandoz-Forschungsinstitut, Brunnerstrasse 59, A-1235 Vienna, Austria

Received 13 September 1994
Accepted 28 December 1994

Keywords: CoMFA; Dihydrofolate reductase inhibitors; Alignment; Indicator variables

Summary

The steric descriptors commonly used in CoMFA – Lennard-Jones 6–12 potential-derived interaction energies calculated between a probe atom and the molecules under investigation – have been replaced by variables indicating the presence of an atom of a particular molecule in predefined volume elements (cubes) within the region enclosing the ensemble of superimposed molecules. The resulting 'atom indicator vectors' were used as steric fields in the subsequent PLS analyses, with and without inclusion of electrostatic Coulomb interaction-derived fields. Application of this method to five training sets (80 compounds each) and five test sets (60 compounds each), randomly selected from an ensemble of 256 dihydrofolate reductase inhibitors, leads to models of significantly higher consistency, as indicated by the cross-validated r^2 values for the training sets and the predictive r^2 values for the test sets.

Introduction

Comparative Molecular Field Analysis (CoMFA) [1] has proven an especially useful QSAR technique in the field of medicinal chemistry, where a number of applications have become known [2–7]. Following definition of a superposition rule for the 3D representations of a set of molecules, the steric and electrostatic interaction energies between a probe atom and each of the structures are calculated at the surrounding points of a predefined grid. From the resulting, highly underdetermined matrices, linear equations are derived using the partial least squares (PLS) algorithm [8–10]. This regression method is usually performed in combination with cross-validation in order to monitor the consistency of the model [1,11], as indicated by the cross-validated r^2 value (r^2_{cv}).

The steric interaction energies are calculated by means of a Lennard-Jones 6–12 potential which is characterized by a very steep increase in energy at short distances [12]. This may lead to significantly different energy values at grid points close to the structures if one compares two identical molecules which are not perfectly superimposed. In combination with the standard grid spacing of 2.0 Å, rather random field values are assigned to some lattice intersections in the vicinity of the compounds. As the highest variance in energy values is found at these points in particular, the results of the PLS analysis become significantly influenced by this random assignment. Therefore, the definition of a very accurate alignment rule for the compounds has proven to be of crucial importance for analyses involving Lennard-Jones interaction energies.

In CoMFA, the 3D structures of the molecules are commonly described by their surrounding Lennard-Jones-derived fields. Another method for generating 3D representations of molecules is the mapping of their atoms onto a predefined grid. Such approaches [13], particularly considering the different van der Waals radii of the atoms [14,15], have been described previously. In these approaches, different values are assigned to grid points, depending on whether a grid point is located within, or outside, the van der Waals radius of any atom of a particular molecule. In our investigation, we did not consider different van der Waals radii, and we mapped all atoms of the compounds directly onto a predefined grid. As the defini-

*To whom correspondence should be addressed.

tion of a grid within a particular region is equivalent to filling that region with little cubes, the atom mapping corresponded to checking each of these cubes for atom occupancy. The presence or absence of an atom in a particular cube was determined solely by the Cartesian coordinates of its center (nucleus). Depending on whether or not an atom of a particular molecule was found in such a cube, different values were assigned to the respective grid points. These values will be referred to as atom-based indicator variables. The resulting vectors were used as steric fields in the analyses and the descriptor matrices were analyzed with the PLS method.

For a validation of this method, the classical QSAR data set of Hansch et al., consisting of 256 dihydrofolate inhibitors [16], was used. Five different sets (140 compounds each) were randomly chosen and divided into training sets (80 structures) and test sets (60 entries).

As we were interested to carry out a thorough comparison with the 'classical' CoMFA method, we did not only compare analyses with the two different steric field types, but additionally included Coulomb-derived fields for the electrostatic properties of the molecules into the analyses. In order to avoid unjustified large parametric variance due to the steep increase of the steric field contribution at lattice points close to the molecules, some authors have suggested a truncation of the steric energies at 4.0 or 5.0 kcal/mol instead of the commonly used 30.0 kcal/mol [17–19]. Therefore, we carried out additional analyses using a cutoff value of 5.0 kcal/mol and compared the results with those obtained using a 30 kcal/mol threshold for the steric energies as well.

Methods

All modelling work was performed with the program SYBYL, version 6.0 or 6.04 [20], run on Silicon Graphics workstations (INDY or INDIGO-2, 64 MB main memory, 128 MB swap space, IRIX 5.2 operating system). Table 1 lists the molecule identities (IDs) contained in the five data sets, which were randomly chosen from the compounds listed in Table 1 of Ref. 16. The IDs used in this paper are identical to the original ones. The structures were built according to the rules depicted in Fig. 1. Energy minimizations were performed with the Tripos standard force field [21], without inclusion of electrostatics, using the POWELL minimization technique [22]. The convergence criterion was defined as an energy change ≤ 0.05 kcal/mol between subsequent minimization steps. Partial charges were calculated with MOPAC 5.0 [23] by means of the MNDO method [24]. The standard deviation threshold for exclusion of columns from the PLS analysis was set to 2.0. Cross-validation was performed by means of the 'leave-one-out' technique. Preliminary analyses were carried out with a maximum of seven components. Subsequently, the analyses were repeated using that number of components at which the difference in the cross-validated r^2 value (r^2_{cv}) to the next one was less than 0.02. This procedure was chosen since the number of

TABLE 1
COMPOUND IDs OF TRAINING AND TEST SETS TAKEN FROM REF. 16

Set	Training set IDs	Test set IDs
1	4, 9, 17, 18, 20, 21, 28, 36, 37, 41, 42, 45, 48, 56, 60, 66, 68, 70, 75, 80, 88, 98, 100, 102, 104, 107, 111, 113, 115, 118, 119, 127, 131, 132, 136, 141, 142, 143, 145, 149, 156, 158, 159, 166, 167, 170, 172, 175, 177, 181, 182, 185, 190, 191, 193, 194, 198, 199, 202, 203, 204, 205, 207, 208, 211, 213, 214, 215, 217, 223, 224, 227, 230, 233, 237, 242, 246, 247, 252, 254	1, 3, 5, 7, 10, 19, 23, 25, 27, 31, 35, 39, 46, 57, 61, 62, 63, 69, 71, 72, 74, 76, 77, 78, 81, 83, 84, 86, 91, 94, 95, 96, 99, 101, 105, 108, 114, 120, 122, 125, 129, 138, 139, 155, 160, 163, 168, 179, 184, 201, 216, 219, 220, 222, 226, 228, 231, 232, 238, 244
2	1, 7, 9, 11, 12, 14, 20, 21, 22, 24, 50, 53, 55, 58, 59, 60, 62, 63, 64, 65, 68, 76, 81, 88, 89, 91, 92, 96, 98, 105, 106, 108, 111, 115, 117, 118, 121, 124, 126, 129, 133, 136, 138, 139, 143, 147, 151, 153, 157, 161, 163, 167, 168, 172, 173, 176, 177, 180, 185, 189, 191, 197, 198, 200, 202, 203, 205, 208, 209, 213, 225, 234, 236, 240, 241, 243, 252, 253, 254, 255	2, 6, 13, 19, 28, 30, 31, 35, 38, 40, 43, 57, 61, 67, 71, 72, 79, 83, 86, 90, 94, 97, 99, 102, 104, 107, 109, 110, 112, 123, 127, 130, 132, 145, 152, 155, 159, 170, 174, 178, 182, 187, 192, 195, 199, 207, 214, 217, 220, 221, 222, 223, 226, 229, 230, 233, 235, 237, 246, 247
3	5, 6, 10, 15, 19, 24, 27, 28, 30, 38, 42, 45, 46, 47, 48, 49, 53, 56, 57, 61, 65, 77, 80, 90, 94, 95, 96, 106, 108, 109, 111, 122, 123, 125, 132, 135, 137, 138, 143, 145, 148, 150, 151, 154, 161, 166, 171, 173, 175, 176, 179, 181, 182, 184, 187, 188, 189, 190, 191, 193, 200, 203, 204, 205, 212, 214, 216, 219, 221, 226, 228, 239, 245, 246, 247, 248, 249, 251, 252, 256	8, 12, 14, 16, 17, 18, 20, 21, 22, 23, 29, 32, 34, 40, 52, 54, 67, 69, 71, 73, 74, 76, 79, 82, 83, 84, 85, 93, 105, 110, 112, 115, 118, 119, 127, 130, 131, 139, 141, 144, 146, 149, 156, 159, 162, 163, 174, 178, 186, 196, 207, 208, 209, 211, 215, 223, 231, 234, 243, 253
4	2, 4, 7, 10, 22, 25, 26, 30, 33, 36, 37, 42, 43, 50, 52, 53, 55, 58, 59, 60, 61, 64, 67, 69, 70, 71, 72, 73, 75, 82, 88, 91, 92, 95, 102, 105, 106, 110, 122, 130, 132, 133, 136, 139, 140, 142, 146, 151, 152, 153, 156, 157, 158, 164, 165, 166, 178, 179, 183, 188, 190, 191, 196, 198, 200, 203, 204, 210, 213, 217, 222, 229, 230, 233, 234, 237, 238, 247, 248, 254	6, 8, 14, 23, 24, 31, 32, 35, 38, 39, 41, 48, 66, 68, 74, 77, 78, 84, 89, 98, 100, 107, 108, 109, 111, 116, 118, 119, 120, 121, 124, 126, 128, 129, 135, 137, 143, 144, 148, 149, 154, 161, 170, 175, 176, 177, 180, 189, 193, 195, 197, 202, 215, 224, 227, 243, 245, 249, 251, 255
5	1, 5, 11, 12, 15, 16, 19, 20, 28, 40, 44, 45, 46, 49, 51, 54, 56, 62, 63, 65, 79, 80, 83, 86, 87, 90, 93, 94, 96, 97, 99, 101, 103, 104, 112, 113, 115, 117, 123, 125, 127, 131, 134, 141, 145, 147, 150, 162, 167, 168, 169, 171, 172, 173, 174, 182, 185, 186, 187, 192, 201, 205, 206, 208, 209, 211, 219, 220, 221, 223, 226, 231, 232, 235, 240, 241, 242, 244, 253, 256	6, 13, 17, 18, 22, 24, 26, 29, 31, 32, 39, 41, 48, 50, 59, 64, 69, 70, 75, 77, 78, 91, 98, 100, 107, 109, 124, 126, 128, 129, 133, 137, 138, 139, 140, 142, 157, 161, 163, 164, 170, 175, 178, 180, 191, 197, 198, 204, 213, 222, 224, 225, 227, 234, 238, 239, 247, 248, 249, 255

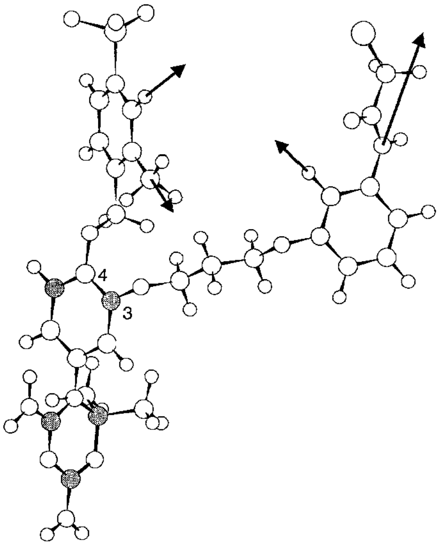

Fig. 1. Representation of the alignment points (shaded grey) and the overall conformation of the dihydrofolate reductase inhibitors. The molecule displayed is a 'hybrid' of two representatives of the major structural modifications in the series (the 3-substituent of compound 137 is linked to structure 100, the compound IDs are identical to those in Ref. 16). Substituents were built in an extended conformation. The arrows indicate the directionality of putative 'non-para' substituents on the phenyl rings in substituents at positions 3 or 4. If the number of chemical groups at a phenyl moiety in 'non-para' position is more than 1, the arrows indicate the position of the largest one.

components at which the initial steep increase of r^2_{cv} starts to level off has proven to give better predictive CoMFA models [25].

The calculation of the predictive r^2 value (r^2_{pred}) was based solely on the molecules in the test set and is defined in analogy to the cross-validated r^2 [1,11]:

$$r^2_{pred} = \frac{SD - PRESS}{SD} \quad (1)$$

where SD is the variance of the biological activities of the molecules in the test set around the mean activity of the training set molecules. 'PRESS' represents the sum of the squared differences between predicted and actual target property values for every compound in the test set. A negative r^2_{pred} reflects a complete lack of predictive ability of the model under consideration.

For the field evaluations involving distance-dependent interaction energies (i.e. Lennard-Jones, Coulomb), the magnitude of the regions was defined to extend the ensemble of superimposed molecules by 4.0 Å along the principal axes of a Cartesian coordinate system (hereafter referred to as grid position A or gpA). An sp^3 carbon with a charge of +1.0 served as a probe atom. For the steric fields, grid spacings of 2.0, 1.0 and 0.75 Å were chosen. At these distances of the lattice intersections, a value of 30.0 kcal/mol was applied as a cutoff for the steric interaction energies. Additional analyses were performed at grid spacings of 2.0 and 1.0 Å, using a threshold of 5.0 kcal/mol. The conditions for evaluation of the electrostatic fields were always the same: the distance of the lattice intersections was 2.0 Å, and the Coulomb interactions were not evaluated at points where the steric interaction energy would exceed the standard cutoff value of 30.0 kcal/mol in the corresponding steric field (i.e., at lattice intersections 'inside' a molecule).

In the atom-mapping procedure, two different grid positions were used. The first one was identical to position A for evaluation of the distance-dependent interaction energies. The second region was defined to extend the superimposed molecules by 0.2 Å (grid position B or gpB). If an atom (i.e its center, nucleus) of a particular molecule was found in one of the cubes defined by the grid, the value at the corresponding lattice intersection was set to 30.0, otherwise to 0.0 (Fig. 2). A value of 30.0 was chosen in order to have an a priori similar scaling of this field type to the 6–12 potential-derived descriptors. This was necessary, since the present implementation of PLS within SYBYL first compares the standard deviation of the columns to a user-specified threshold ('minimum sigma') in order to exclude those columns with little or no variance. Subsequently, the CoMFA standard scaling is performed [26]. Four different lattice spacings, of 2.0, 1.0, 0.75 and 0.57 Å, were investigated. A grid spacing of 0.57 Å corresponds to a spatial diagonal of 0.99 Å for the respective cubes, thus excluding the possibility to find two atoms of a particular molecule within the same cube.

Results and Discussion

In Tables 2 and 3, the statistical parameters of the CoMFAs are summarized. By comparing the different analyses, some general trends can be derived: (i) In the CoMFAs with the standard 6–12 potentials, a reduction of the grid spacing does not lead to an improvement of the statistical parameters (r^2_{cv} and r^2_{pred}). (ii) In contrast, for the analyses using indicator fields, narrower lattice spa-

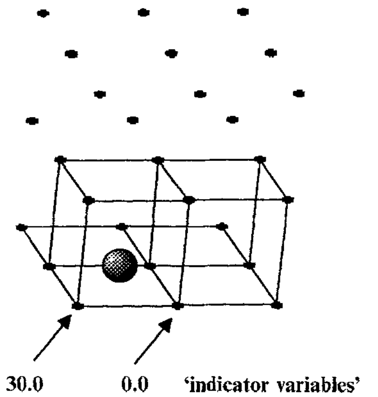

Fig. 2. Schematic representation of the atom-mapping procedure. The sphere represents an atom center. If an atom (i.e. its center) is found within one of the cubes, a value of 30.0 is assigned to the corresponding grid point.

TABLE 2
STATISTICAL PARAMETERS OF THE CoMFAs INCLUDING STERIC FIELDS

Set	Grid position	Steric field type[a]	Minimum sigma	Grid (Å)	Steric cutoff	No. of columns used	r^2_{cv}[b]	r^2_{pred}
1	A	LJ	2.00	2.00	30.0	319	0.160 (3)	0.450
	A	LJ	2.00	2.00	5.0	89	0.007 (2)	0.282
	A	LJ	0.50	2.00	5.0	351	0.145 (2)	0.407
	B	I	2.00	2.00	–	197	0.193 (2)	0.420
	A	I	2.00	2.00	–	193	0.168 (2)	0.479
	A	LJ	2.00	1.00	30.0	2556	0.148 (3)	0.463
	A	LJ	0.50	1.00	5.0	2773	0.137 (2)	0.409
	B	I	2.00	1.00	–	566	0.177 (2)	0.581
	A	I	2.00	1.00	–	574	0.262 (2)	0.586
	A	LJ	2.00	0.75	30.0	6047	0.138 (3)	0.453
	B	I	2.00	0.75	–	760	0.334 (2)	0.595
	A	I	2.00	0.75	–	763	0.316 (4)	0.632
	B	I	2.00	0.57	–	933	0.265 (2)	0.629
	A	I	2.00	0.57	–	935	0.150 (2)	0.669
2	A	LJ	2.00	2.00	30.0	316	0.534 (3)	0.214
	A	LJ	2.00	2.00	5.0	82	0.378 (3)	0.063
	A	LJ	0.50	2.00	5.0	361	0.503 (3)	0.194
	B	I	2.00	2.00	–	199	0.422 (3)	0.214
	A	I	2.00	2.00	–	201	0.459 (3)	0.224
	A	LJ	2.00	1.00	30.0	2550	0.526 (3)	0.225
	A	LJ	0.50	1.00	5.0	2780	0.529 (3)	0.222
	B	I	2.00	1.00	–	584	0.451 (4)	0.516
	A	I	2.00	1.00	–	588	0.549 (3)	0.415
	A	LJ	2.00	0.75	30.0	6045	0.522 (3)	0.221
	B	I	2.00	0.75	–	786	0.547 (3)	0.487
	A	I	2.00	0.75	–	810	0.546 (3)	0.545
	B	I	2.00	0.57	–	918	0.594 (3)	0.517
	A	I	2.00	0.57	–	944	0.510 (4)	0.424
3	A	LJ	2.00	2.00	30.0	348	0.286 (3)	0.272
	A	LJ	2.00	2.00	5.0	87	0.207 (3)	-0.054
	A	LJ	0.50	2.00	5.0	397	0.305 (3)	0.238
	B	I	2.00	2.00	–	224	0.251 (3)	0.343
	A	I	2.00	2.00	–	201	0.459 (3)	0.224
	A	LJ	2.00	1.00	30.0	2889	0.271 (3)	0.248
	A	LJ	0.50	1.00	5.0	3151	0.331 (5)	0.386
	B	I	2.00	1.00	–	618	0.412 (3)	0.425
	A	I	2.00	1.00	–	588	0.549 (3)	0.415
	A	LJ	2.00	0.75	30.0	6834	0.265 (3)	0.245
	B	I	2.00	0.75	–	827	0.528 (5)	0.565
	A	I	2.00	0.75	–	810	0.543 (4)	0.466
	B	I	2.00	0.57	–	993	0.506 (3)	0.420
	A	I	2.00	0.57	–	970	0.476 (4)	0.408
4	A	LJ	2.00	2.00	30.0	301	0.359 (3)	0.535
	A	LJ	2.00	2.00	5.0	71	0.113 (1)	0.151
	A	LJ	0.50	2.00	5.0	338	0.342 (3)	0.520
	B	I	2.00	2.00	–	180	0.332 (2)	0.593
	A	I	2.00	2.00	–	187	0.308 (2)	0.650
	A	LJ	2.00	1.00	30.0	2389	0.344 (3)	0.603
	A	LJ	0.50	1.00	5.0	2599	0.340 (3)	0.598
	B	I	2.00	1.00	–	536	0.470 (2)	0.619
	A	I	2.00	1.00	–	551	0.485 (4)	0.707
	A	LJ	2.00	0.75	30.0	5607	0.336 (3)	0.600
	B	I	2.00	0.75	–	730	0.450 (2)	0.666
	A	I	2.00	0.75	–	745	0.574 (3)	0.665
	B	I	2.00	0.57	–	897	0.402 (2)	0.602
	A	I	2.00	0.57	–	902	0.462 (3)	0.614
5	A	LJ	2.00	2.00	30.0	345	0.618 (5)	0.478
	A	LJ	2.00	2.00	5.0	82	0.451 (4)	0.171
	A	LJ	0.50	2.00	5.0	370	0.633 (5)	0.506
	B	I	2.00	2.00	–	212	0.445 (3)	0.455
	A	I	2.00	2.00	–	216	0.668 (4)	0.498
	A	LJ	2.00	1.00	30.0	2785	0.587 (4)	0.427

TABLE 2 (continued)

Set	Grid position	Steric field type[a]	Minimum sigma	Grid (Å)	Steric cutoff	No. of columns used	r^2_{cv} [b]	r^2_{pred}
5	A	LJ	0.50	1.00	5.0	3024	0.598 (4)	0.438
	B	I	2.00	1.00	–	601	0.547 (4)	0.494
	A	I	2.00	1.00	–	616	0.639 (3)	0.569
	A	LJ	2.00	0.75	30.0	6593	0.589 (4)	0.431
	B	I	2.00	0.75	–	808	0.659 (3)	0.620
	A	I	2.00	0.75	–	808	0.578 (3)	0.559
	B	I	2.00	0.57	–	951	0.613 (3)	0.438
	A	I	2.00	0.57	–	943	0.499 (3)	0.557

[a] LJ = Lennard-Jones; I = Indicator.

[b] Numbers in parentheses indicate the number of components used in PLS analysis.

cings result in a significant increase of the r^2_{cv} and r^2_{pred} values. (iii) The attempt to improve the standard CoMFAs by truncating the probe–ligand steric energies at a value lower than the default setting (5.0 instead of 30.0) does not yield significant improvements. (iv) Comparing the results obtained with the two different steric field types after inclusion of electrostatic descriptors in terms of Coulomb-derived interaction energies, the analyses with the indicator fields are still superior. (v) The analyses with indicator fields show in some cases a significant dependency on the grid position used. However, at both positions investigated they are superior to those using Lennard-Jones-derived fields.

In the following, we present and discuss the results more thoroughly; comparisons with similar approaches will be performed as well.

Different grid spacings

In the analyses with a grid spacing of 2.0 Å, the Lennard-Jones-derived steric fields gave better results than the indicator fields with respect to some of the training sets (sets 2, 3, 4 and 5 at gpB; sets 2 and 4 at gpA; Table 1); however, for some of the test sets (sets 3 and 4 at gpB; sets 1, 2, 4 and 5 at gpA) the indicator fields were superior. Evaluation of the indicator fields at a lattice spacing of 1.0 Å led to a significant increase of both the r^2_{cv} and the r^2_{pred} values. With respect to the r^2_{pred}, the use of indicator fields was generally superior to the use of Lennard-Jones potentials: in some analyses (sets 1–3 at gpB; all sets at gpA) a remarkable increase in r^2_{pred} was apparent. This trend was continued at a spacing between the lattice intersections of 0.75 Å; only in set 5 at gpA were the indicator fields slightly worse with respect to r^2_{cv}. With respect to the test sets (r^2_{pred}), the CoMFAs using indicator fields were superior in all cases. For most of the sets, both r^2 values of the analyses with indicator fields decreased at the finest grid (0.57 Å) compared to the CoMFAs with this field type at 0.75 Å.

The Lennard-Jones-derived steric interaction energies with a cutoff value of 30.0 kcal/mol were evaluated at grid spacings of 2.0, 1.0 and 0.75 Å. In the corresponding analyses, distances smaller than 2.0 Å between the lattice intersections led to a slight decrease of the cross-validated r^2 values. This result is in fact not surprising, as it is known that a reduction of the lattice spacing does not improve r^2_{cv} [1,11,26–28,30]; most of the associated increase in field information is noise insofar as a PLS correlation is concerned. For the r^2_{pred}, in two cases (sets 1 and 2) the results remained approximately the same, for two sets (3 and 5) they became worse, and in one case (set 4) an improvement of r^2_{pred} was apparent.

With the Lennard-Jones fields, an exponential increase in the number of columns used at narrower grid spacings became apparent. At a grid spacing of 0.57 Å, we were not able to perform analyses with this field type; the computational effort exceeded the ability of our workstations due to memory overload. On the other hand, for the analyses with indicator fields, the increase in the number of columns used at narrower grid spacings was much less prominent. Whereas at a distance of the lattice intersections of 0.75 Å, 6000–7000 columns were used in the analyses with Lennard-Jones-derived steric fields, only 700–800 columns were selected in the CoMFAs using indicator fields.

The results after inclusion of electrostatic descriptors in terms of Coulomb interaction energies are listed in Table 3. The overall differences between both r^2 values of the analyses with the two different steric descriptor types were slightly reduced compared to the models derived with steric descriptors only. With respect to the r^2_{cv} values for set 2, the CoMFAs including Lennard-Jones-derived steric fields yielded 'globally' slightly higher values than the analyses with indicator fields at both grid positions. Nevertheless, in those cases where a rather low r^2_{cv} was obtained by using the Lennard-Jones steric fields, the indicator fields still yielded significantly higher values (sets 1 and 4). The r^2_{pred} values were in almost all cases higher using indicator fields (exceptions: set 1: 2.0 Å; set 5: 2.0 Å at gpB); the best r^2_{pred} was always obtained with the indicator fields.

On average, for the analyses using indicator fields, the grid spacing of 0.75 Å gave the best results. In many cases, a decrease of the statistical parameters became apparent at a narrower distance of the lattice intersections (0.57 Å). This phenomenon may be interpreted as a compromise of two contrary developments: on the one hand,

TABLE 3
STATISTICAL PARAMETERS OF THE CoMFAs INCLUDING STERIC AND ELECTROSTATIC FIELDS

Set	Grid position	Steric field type[a]	Minimum sigma	Grid (Å)	Steric cutoff	Steric contribution	No. of columns used	r^2_{cv} [b]	r^2_{pred}
1	A	LJ	2.00	2.00	30.0	0.50	497	0.205 (3)	0.385
	A	LJ	2.00	2.00	5.0	0.50	267	0.097 (3)	0.316
	A	LJ	0.50	2.00	5.0	0.44	1022	0.196 (3)	0.403
	B	I	2.00	2.00	–	0.52	375	0.206 (3)	0.371
	A	I	2.00	2.00	–	0.52	371	0.237 (4)	0.519
	A	LJ	2.00	1.00	30.0	0.50	2734	0.204 (3)	0.393
	A	LJ	0.50	1.00	5.0	0.57	3444	0.204 (3)	0.401
	B	I	2.00	1.00	–	0.55	744	0.256 (4)	0.559
	A	I	2.00	1.00	–	0.54	752	0.308 (4)	0.564
	A	LJ	2.00	0.75	30.0	0.50	6225	0.197 (3)	0.386
	B	I	2.00	0.75	–	0.54	938	0.356 (4)	0.530
	A	I	2.00	0.75	–	0.54	941	0.304 (4)	0.544
	B	I	2.00	0.57	–	0.52	1111	0.329 (4)	0.567
	A	I	2.00	0.57	–	0.51	1113	0.214 (4)	0.592
2	A	LJ	2.00	2.00	30.0	0.57	489	0.596 (4)	0.286
	A	LJ	2.00	2.00	5.0	0.40	255	0.460 (3)	0.099
	A	LJ	0.50	2.00	5.0	0.39	1031	0.582 (4)	0.255
	B	I	2.00	2.00	–	0.56	372	0.530 (4)	0.290
	A	I	2.00	2.00	–	0.57	374	0.505 (3)	0.323
	A	LJ	2.00	1.00	30.0	0.57	2723	0.591 (4)	0.294
	A	LJ	0.50	1.00	5.0	0.61	3450	0.593 (4)	0.296
	B	I	2.00	1.00	–	0.57	757	0.517 (4)	0.420
	A	I	2.00	1.00	–	0.58	761	0.566 (4)	0.382
	A	LJ	2.00	0.75	30.0	0.57	6218	0.588 (4)	0.284
	B	I	2.00	0.75	–	0.57	959	0.545 (4)	0.439
	A	I	2.00	0.75	–	0.57	983	0.534 (4)	0.463
	B	I	2.00	0.57	–	0.57	1091	0.540 (4)	0.505
	A	I	2.00	0.57	–	0.55	1117	0.461 (4)	0.355
3	A	LJ	2.00	2.00	30.0	0.48	526	0.405 (4)	0.311
	A	LJ	2.00	2.00	5.0	0.51	265	0.325 (4)	0.124
	A	LJ	0.50	2.00	5.0	0.47	1113	0.404 (4)	0.362
	B	I	2.00	2.00	–	0.49	402	0.384 (4)	0.335
	A	I	2.00	2.00	–	0.47	403	0.381 (4)	0.331
	A	LJ	2.00	1.00	30.0	0.47	3067	0.371 (4)	0.311
	A	LJ	0.50	1.00	5.0	0.52	3867	0.360 (4)	0.365
	B	I	2.00	1.00	–	0.50	796	0.443 (3)	0.321
	A	I	2.00	1.00	–	0.51	805	0.405 (3)	0.354
	A	LJ	2.00	0.75	30.0	0.47	7012	0.369 (4)	0.309
	B	I	2.00	0.75	–	0.49	1005	0.449 (3)	0.421
	A	I	2.00	0.75	–	0.51	994	0.430 (3)	0.325
	B	I	2.00	0.57	–	0.47	1171	0.470 (3)	0.334
	A	I	2.00	0.57	–	0.48	1148	0.436 (3)	0.299
4	A	LJ	2.00	2.00	30.0	0.44	467	0.334 (3)	0.586
	A	LJ	2.00	2.00	5.0	0.41	237	0.146 (1)	0.184
	A	LJ	0.50	2.00	5.0	0.42	981	0.340 (3)	0.592
	B	I	2.00	2.00	–	0.56	346	0.377 (4)	0.619
	A	I	2.00	2.00	–	0.56	353	0.325 (3)	0.656
	A	LJ	2.00	1.00	30.0	0.55	2555	0.308 (3)	0.630
	A	LJ	0.50	1.00	5.0	0.58	3242	0.326 (3)	0.625
	B	I	2.00	1.00	–	0.62	702	0.454 (4)	0.655
	A	I	2.00	1.00	–	0.60	717	0.466 (5)	0.714
	A	LJ	2.00	0.75	30.0	0.55	5773	0.301 (3)	0.629
	B	I	2.00	0.75	–	0.57	896	0.401 (4)	0.683
	A	I	2.00	0.75	–	0.63	911	0.557 (5)	0.680
	B	I	2.00	0.57	–	0.54	1063	0.429 (3)	0.629
	A	I	2.00	0.57	–	0.59	1068	0.466 (5)	0.662
5	A	LJ	2.00	2.00	30.0	0.60	523	0.586 (4)	0.426
	A	LJ	2.00	2.00	5.0	0.41	260	0.488 (3)	0.271
	A	LJ	0.50	2.00	5.0	0.37	1090	0.634 (5)	0.463
	B	I	2.00	2.00	–	0.60	390	0.481 (3)	0.417
	A	I	2.00	2.00	–	0.63	394	0.588 (4)	0.513
	A	LJ	2.00	1.00	30.0	0.60	2963	0.595 (4)	0.440

TABLE 3 (continued)

Set	Grid position	Steric field type[a]	Minimum sigma	Grid (Å)	Steric cutoff	Steric contribution	No. of columns used	r^2_{cv} [b]	r^2_{pred}
5	A	LJ	0.50	1.00	5.0	0.64	3744	0.606 (4)	0.450
	B	I	2.00	1.00	–	0.63	779	0.547 (4)	0.467
	A	I	2.00	1.00	–	0.61	794	0.609 (5)	0.527
	A	LJ	2.00	0.75	30.0	0.60	6771	0.589 (4)	0.431
	B	I	2.00	0.75	–	0.60	986	0.613 (5)	0.545
	A	I	2.00	0.75	–	0.62	986	0.519 (5)	0.491
	B	I	2.00	0.57	–	0.60	1129	0.557 (5)	0.458
	A	I	2.00	0.57	–	0.58	1121	0.439 (4)	0.501

[a] LJ = Lennard-Jones; I = Indicator.

[b] Numbers in parentheses indicate the number of components used in PLS analysis.

the shape of the structures should be described exactly. On the other hand, the degree of differentiation should not be too high. Atoms of different molecules which are located at almost identical positions in space should be described as being equal. A very fine grid will differentiate such atoms and puts the corresponding indicator values into different columns of the descriptor matrix, thus describing these two atoms as being not superimposable. However, this is not the intention of the method, since it should level out large differences in the descriptors for 'similar' atoms. Therefore, the grid spacing of 0.75 Å appears to be the best compromise between exactness of shape description and inaccuracy in differentiation of atoms.

Truncation of the probe–ligand steric energies

It has been suggested that a truncation of the probe–ligand steric energies in CoMFA to 4.0 or 5.0 kcal/mol might lead to analyses of higher consistency, as an unjustified large parametric variance at lattice points close to the molecules should be avoided [17–19]. Therefore, analyses with a cutoff value of 5.0 kcal/mol were performed in this study as well. With the standard deviation threshold ('minimum sigma') at 2.0 kcal/mol in order to exclude columns with little or no variance from the PLS analyses, both r^2 values (r^2_{cv} and r^2_{pred}) became worse in all cases. Nevertheless, this reduction of the statistical parameters was related to an overproportional exclusion of columns; apparently the relatively few remaining ones did not contain enough information. After setting the standard deviation threshold to 0.5 kcal/mol, approximately the same number of columns remained in the PLS analyses, leading to similar results compared to the standard method (minimum sigma: 2.0, steric cutoff: 30.0). In contrast to results described by Klebe and Abraham [19], a reduction of the steric cutoff value did not lead to superior analyses. Only for set 3 in Table 1, an improvement of the CoMFAs was detectable in our study. A possible explanation for this discrepancy might be the rather limited size of the sets investigated (8, 20 and 13 compounds) in the above-mentioned paper. With data sets of this size, a different degree of accuracy in the prediction of a single compound can already lead to a significantly different r^2_{cv}. Support for this hypothesis is also given by the fact that for the first data set investigated (eight molecules, HRV14, Table 4 in the original reference), one of the two r^2_{cv} values reported decreases significantly. However, an exact comparison with these results is difficult because no standard deviation threshold values were reported for the analyses. Unfortunately, the other authors cited [17,18] did not perform any comparisons by using different truncation values for the steric energies.

Steric versus electrostatic contributions

The steric and electrostatic contributions were generally balanced, the extrema being approximately 60% in favour of one of these field types. At a grid spacing of 2.0 Å, truncation of the steric fields to a maximum of 5.0 kcal/mol and application of a standard deviation threshold of 0.5 kcal/mol led to a preference of the electrostatic contribution relative to the standard method (30/30 cutoff, minimum sigma 2.0). In these cases, the different truncation values (5.0 for the steric, 30.0 for the electrostatic contribution), in combination with a rather low minimum sigma, gave rise to a higher proportion of electrostatic columns, resulting in a preference of the electrostatic contribution.

Comparison to similar approaches

Other approaches to the shape description of molecules in combination with PLS have been published. Wiese and Coats [29] modified the HASL method [15] and combined it with PLS. They investigated two different training sets (10 compounds each) and reported a more accurate prediction for test sets of five and four molecules, respectively, compared to the original HASL approach. Nevertheless, a direct comparison of our results to their investigation is not possible. First, they did not perform comparisons with CoMFA. Second, a fundamental difference in the calculation of their steric fields is the fact that van der Waals radii were considered. Whether they included electrostatics in their PLS analyses is not clarified in their publication. Also no information on the grid spacing applied is provided; however, within the HASL approach distances of the lattice intersections in the range of 2.8–4.0 Å appear to be usual.

In their study of 'shape potentials' in combination with PLS, Floersheim et al. computed steric fields in a very similar manner [30]. They assigned values of either 1 or 0 to grid points, depending on whether the grid point is within, or outside, the van der Waals radius of any atom of the molecule in a predefined grid (distance of the lattice intersections: 2.0 Å) [14]. A data set of 19 compounds was analyzed using their own suite of computer programs called COMPA (Comparison of Molecular Potentials and Analysis). They compared results obtained by using the shape potentials with the standard CoMFA approach. At a grid spacing of 2.0 Å, the r^2_{cv} values for the two different methods (for the Lennard-Jones-derived steric fields a hydrogen was used as a probe atom) were rather similar. Results on the prediction of test compounds were not reported. Remarkably, at finer grids (1.5, 1.0 and 0.75 Å) the r^2_{cv} decreased, which may be related to the basic difference in calculation of these shape potentials compared to our approach. As we did not consider the van der Waals radii of the atoms, the finer grid led to a higher resolution and consequently to an improvement of the results compared to the classical CoMFA method.

Conclusions

The replacement of steric Lennard-Jones 6–12 derived interaction energies by atom-based indicator variables appears to be a promising method to further enhance the consistency and the predictiveness of CoMFA models. This method provides a less arbitrary assignment of values indicating the presence of atoms at specific locations in space, thus leading to higher r^2 values compared with the standard method. In order to achieve improvements of the r^2 values, a rather fine grid must be applied. In contrast to the analyses with Lennard-Jones potentials, small grid spacings do not lead to an exponential increase of the columns used. Apparently, in this type of analyses the reduction of the lattice spacing leads to a higher information content without increasing the noise.

As also observed in classical CoMFA studies, the positioning of the grid has an influence on the statistical parameters of the respective analyses using indicator fields. Nevertheless, at the two grid positions investigated, the atom-mapping procedure turned out to give better results compared to the standard CoMFAs.

Future investigations should also include the treatment of structures containing large rings. Up to now, no differentiation between 'inside' and 'outside' of the molecules has been made.

Acknowledgements

The authors would like to thank Peter Stütz and Heinz Berner for reading the manuscript and many helpful discussions.

References

1 Cramer III, R.D., Patterson, D.E. and Bunce, J.E., J. Am. Chem. Soc., 110 (1988) 5959.
For applications of CoMFA in medicinal chemistry, see for instance Refs. 2–7:

2 Avery, M.A., Gao, F. and Chong, W.K.M., J. Med. Chem., 36 (1993) 4264.

3 Horwitz, J.P., Massova, I., Wiese, T.E., Besler, B.H. and Corbett, T.H., J. Med. Chem., 37 (1994) 781.

4 Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R., J. Med. Chem., 36 (1993) 4152.

5 Waller, C.L. and Marshall, G.R., J. Med. Chem., 36 (1993) 2390.

6 DePriest, S.A., Mayer, D., Naylor, C.B. and Marshall, G.R., J. Am. Chem. Soc., 115 (1993) 5372.

7 Debnath, A.K., Hansch, C., Kim, K.H. and Martin, Y.C., J. Med. Chem., 36 (1993) 1007.

8 Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J., SIAM J. Sci. Stat. Comput., 5 (1984) 735.

9 Wold, S., Albano, C., Dunn III, W.J., Edlund, U., Esbenson, K., Geladi, P., Hellberg, S., Johannson, E., Lindberg, W. and Sjörström, M., In Kowalski, B. (Ed.) Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, 1984, pp. 17–95.

10 Stahle, L. and Wold, S., Prog. Med. Chem., 25 (1988) 292.

11 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., Quant. Struct.–Act. Relatsh., 7 (1988) 18.

12 Thibaut, U., Folkers, G., Klebe, G., Kubinyi, H., Merz, A. and Rognan, D., Quant. Struct.–Act. Relatsh., 13 (1994) 1.

13 Snell, C., (1994) personal communication.

14 Marsili, M., Floersheim, P. and Dreiding, A.S., Comput. Chem., 7 (1983) 175.

15 Doweyko, A.M., J. Med. Chem., 31 (1988) 1396.

16 Silipo, C. and Hansch, C., J. Am. Chem. Soc., 97 (1975) 6849.

17 Kim, K.H. and Martin, Y.C., J. Med. Chem., 34 (1991) 2056.

18 Greco, G., Novellino, E., Silipo, C. and Vittoria, A., Quant. Struct.–Act. Relatsh., 10 (1991) 289.

19 Klebe, G. and Abraham, U., J. Med. Chem., 36 (1993) 70.

20 SYBYL Molecular Modelling Package, Version 6.04, TRIPOS Associates, St. Louis, MO, 1993.

21 Vinter, J.G., Davies, A. and Saunder, M.R., J. Comput.-Aided Mol. Design, 1 (1987) 31.

22 Powell, M.J.D., Math. Program., 12 (1977) 241.

23 Stewart, J.J.P. and Seiler, F.J., MOPAC (Version 5.00), QCPE Program No. 455, Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, 1989.

24 Dewar, M.J.S. and Thiel, W., J. Am. Chem. Soc., 99 (1977) 4899.

25 SYBYL Molecular Modelling Software, Version 6.0 Theory Manual, Tripos Associates, St. Louis, MO, 1992, p. 2225.

26 Cramer III, R.D., DePriest, S.A., Patterson, D.E. and Hecht, P., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, p. 465.

27 Calder, J.A., Wyatt, J.A., Frenkel, D.A. and Casida, J.E., J. Comput.-Aided Mol. Design, 7 (1993) 45.

28 Rault, S., Bureau, R., Pilo, J.C. and Robba, M., J. Comput.-Aided Mol. Design, 6 (1992) 553.

29 Wiese, M. and Coats, E.A., Pharmacochem. Libr., 16 (1991) 343.

30 Floersheim, P., Nouzlak, J. and Weber, H.P., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings of the 9th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, p. 227.