MIMUMBA Revisited: Torsion Angle Rules for Conformer Generation Derived from X-ray Structures[†]

Jens Sadowski* and Jonas Boström

Lead Generation Department, AstraZeneca R&D, S-43183 Mölndal, Sweden

Received February 3, 2006

A method has been developed which automatically generates SMARTS patterns for four-atomic torsional fragments, searches experimental structures in the Cambridge Crystallographic Database, and obtains rules for preferred torsion angles in drug-size molecules. These rules can be used for exhaustive conformational analysis using the popular conformer generator OMEGA. This approach results in an overall improvement of quality and coverage of conformational space when comparing conformer ensembles generated by this method with results obtained by using the default OMEGA setup. In particular, the percentage of structures with at least one conformation closer than 0.5 Å to the X-ray structure improves from 84% to 92% in a test set of 11 027 experimental structures from the CSD. Moreover, the average RMS distance of the closest conformation to the X-ray structure improves from 0.30 to 0.22 Å.

INTRODUCTION

Many state-of-the-art molecular modeling methods as, e.g., 3D-database searching or protein−ligand docking, address the flexibility of small molecules by using multiple discrete conformations. Implicitly, multiple conformations are generated on the fly. For example, popular docking programs such as FlexX[1] or GOLD[2] apply lists of reasonable torsion angle values during the build-up or search process. Explicitly, multiple conformations can be generated in advance and be read from databases by the application program. Examples are pharmacophore searches using Catalyst[3] or shape-based searches using programs such as ROCS[4] (small molecule alignment) or FRED[4] (shape-based docking). There are a number of methods available for generating multiple conformers for druglike molecules (for a recent comparison see ref 5). All these conformer generators rely more or less on theoretical assumptions of conformational energy, empiric rules, or both. The underlying knowledge comes from theory (e.g., quantum chemistry, force fields) or experiments (e.g., X-ray crystallography). An elegant way to join experiment and theory was demonstrated in the MIMUMBA program.[6] There, based on the frequency of torsion angles for various four-atomic patterns in experimental structures in the Cambridge Crystallographic Database (CSD)[7] and reverse Boltzmann statistics, torsion angle rules and force field potentials were derived.

The conformer generator OMEGA[4] is widely used as an efficient and fast way to generate multiple conformations of druglike molecules. OMEGA searches torsion angle space by assigning torsion angles to rotatable bonds via SMARTS[8] substructure patterns. The default torsion data file ("torlib.txt") is a handcrafted set of 123 general SMARTS patterns. Since many specific patterns are missing, we and many of our colleagues have tried to improve this set manually but with limited success. Here, we introduce an automatic procedure to obtain many more and more relevant SMARTS patterns from experimental structures in the CSD applying the principles published in the original MIMUMBA paper.[6]

MATERIALS AND METHODS

Overall Strategy. OMEGA uses SMARTS patterns to describe four-atomic torsional fragments and to assign allowed torsion angle values from a database to the corresponding bonds. All angle values are treated equally at this step. That is, there is no discrimination according to energy, etc. Note however that OMEGA uses its internal force field in a subsequent step in order to rank conformations. The overall aim of this work is the automatic generation of torsion data for OMEGA from experimental structures as summarized in the following steps: (1) retrieve a data set of drug-sized organic molecules from the CSD, (2) generate torsion SMARTS automatically, (3) search and analyze the CSD, (4) extract new torsion values, (5) assess performance using crystal structures, and (6) compare rules ("torlib.txt") against force-field optimization (MMFF).

CSD Data Set. A high-quality subset of drug-sized organic compounds was selected from the CSD[7] by using the ConQuest program[9] and in-house software according to the following criteria: (1) molecules labeled as "organic", (2) no error flags set in the database records, (3) no disorders, (4) R-factor < 5%, (5) molecular weight between 100 and 750, (6) maximum six rotatable bonds, and (7) maximum ring size of nine atoms. In addition, counterions and other small species (e.g., solvent) were removed from the individual records as well as duplicate molecules from the database. This procedure gave a total of 31 027 compounds that were randomly split into a training set and a test set of 20 000 and 11 027 compounds, respectively.

Atom Typing and SMARTS. The original MIMUMBA approach[6] uses SYBYL atom types[10] to describe torsional

[†] Dedicated to Professor Johann Gasteiger.

* Corresponding author phone: +46 31 7762850; e-mail: jens.sadowski@astrazeneca.com.

10.1021/ci060042s CCC: $33.50
Published on Web 07/12/2006

fragments. For example, the central bond in butane is described by

"CANY 1 C3(H)(H) 1 C3(H)(H) CANY"

Note that CANY stands for any carbon and C3 for the C.3 atom type and that the 1's denote single bonds. This has to be translated for OMEGA into the corresponding SMARTS pattern [#6][C∧3H2]-!@[C∧3H2][#6] (note that the "∧3" atomic property is a SMARTS extension for hybridization meaning sp^3 in this case). The original rules in MIMUMBA as well as in OMEGA were handcrafted. To automate the generation of all reasonable SMARTS patterns for OMEGA, we followed the procedure published in the MIMUMBA paper:[6] Starting from a table of SYBYL atom types for the different positions in a four-atomic torsional fragment and a set of rules for compatibility and symmetry, all possible torsional patterns are automatically generated. To achieve this, the 49 SYBYL atom types in the MIMUMBA paper were one-to-one translated into corresponding SMARTS patterns and then combined according to the published rules. Note that the MIMUMBA atom types in contrast to the original set of types from SYBYL include also composite types as C3(H)(H) and generic types such as CANY, R, and X. Note also that this encoding scheme does not consider asymmetry. Thus, during torsion angle retrieval there was no control over the absolute configuration of chiral torsion patterns, and all derived torsion rules are consequently symmetrized.

An in-house program based on the OELib[4] C++ toolkit performed the automatic SMARTS generation. This resulted in a total of 52 730 unique and valid SMARTS.

SMARTS Searching and Torsion Rule Derivation. The automatically derived SMARTS were used to search the CSD data set described above to obtain frequency statistics for discrete torsion angle values. Again, an in-house program based on OELib was used. For each SMARTS pattern, the CSD training set of 20 000 X-ray structures was searched, and the total number of hits was stored as well as the number of hits in individual 30-degree bins. The bins for asymmetric torsions run from 0 to 180 degrees and for symmetric patterns from 0 to 90 or from 0 to 120 degrees depending on the symmetry. We used 30-degree bins in order to follow the original OMEGA philosophy—there is no limitation in our approach to use different bin sizes. After obtaining the binned frequencies (or histograms), rules were established to select appropriate torsion angles. This was performed according to the principle that highly occupied torsion angle intervals are more likely than intervals with lower frequency. Since OMEGA unlike MIMUMBA only uses lists of allowed torsion values without further discriminating between them, we settled for two cutoffs: The total number of hits found for an individual SMARTS pattern (hits/SMARTS) and the normalized frequency of an individual torsion angle interval (hits in interval/total hits for pattern) in percent. The first criterion removes SMARTS from the list of torsion patterns if they are under-represented in the training set. The second criterion removes torsion angles from individual torsion rules if the corresponding interval is underpopulated. The two examples in Figure 1 illustrate this. Figure 1 shows torsion angle histograms for butane and ortho-ethylphenol obtained when searching the corresponding SMARTS in the CSD training set. In total 618 hits are returned in the butane case with an average of 59 hits per 30-degree interval, as indicated by a horizontal line in the plot. This corresponds to a value of 100% of the frequency cutoff introduced above and results in a list of allowed torsion angles of ±60 and 180 degrees. Moving the cutoff up or down can change the selection of torsion angles. For the second example in Figure 1 177 hits are obtained from the training set with an average of 26. This results in torsion angles of ±60, ±90, and 180 when setting the frequency cutoff at 100%.

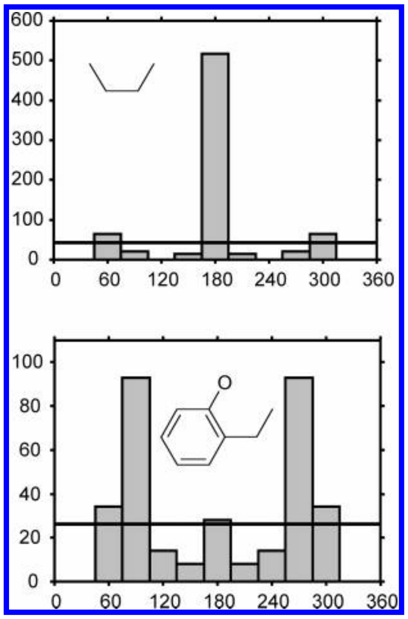

Figure 1. Torsion angle histograms for butane and ortho-ethylphenol. X-axes: torsion angle intervals. Y-axes: number of hits. Horizontal lines indicate a cutoff at the average population.

The final values chosen for hits/SMARTS and the frequency cutoff were derived based on a systematic scan and quantity-quality considerations. This is described in the results section.

OMEGA Performance Measures. To assess the quality of conformer generation based on different sets of torsion rules ("torlib.txt"), OMEGA 1.8.1 was run over the test set of 11 027 crystal structures under the following conditions: (1) max. 256 conformations, (2) energy cutoff: 8 kcal/mol, (3) min. 0.3 Å RMS distance between individual conformers, (4) start from X-ray conformation, and (5) no generation of ring conformations (empty ring conformations file "ringlib.txt"). The two last criteria—starting from X-ray structures and not generating ring conformations—might seem somewhat artificial. We made this choice in order to separate the effect of the torsion rules from other influences. In production runs, we would of course normally not have X-ray structures to start with, and we recommend certain additional procedures (including ring conformations) and parameters to ensure the quality of the results. For a comprehensive study on this topic see ref 5. To quantify and compare the performance of OMEGA with different sets of torsion rules, the following three criteria were defined: (1) coverage—percentage [%] of structures with at least one conformation closer than 0.5 Å to the X-ray structure, (2) average RMS [Å] of the closest OMEGA conformation (averaged over the whole test set), and (3) average number of conformations. The two first criteria monitor the coverage of conformational

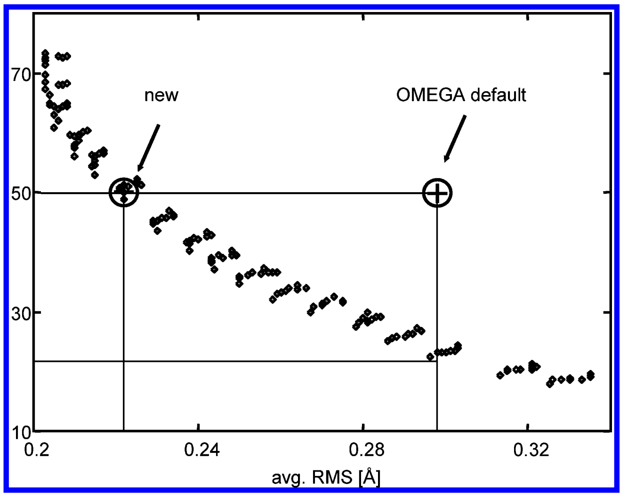

Figure 2. Plot of the number of conformations vs the average RMS of the closest OMEGA conformation compared with the X-ray structure. The small circles represent the 170 experiments for different pairs of hits/SMARTS and frequency cutoffs. The large circles mark the results obtained with the original OMEGA torsion file and with the one based on the chosen parameters for the new torsion file (minimum 20 hits per SMARTS and minimum 45% frequency for each individual angle interval).

space assuming that the X-ray conformation is at least a realistic conformation that should be found by OMEGA.

RESULTS

CSD Searching and Torsion Angle Statistics. The set of 52 730 automatically generated SMARTS was used to search the training set of 20 000 crystal structures from CSD in order to analyze torsion angle distributions. For each SMARTS, the total number of hits and the number of hits in 30-degree bins were reported. After applying an initial cutoff, statistics were obtained for 2863 torsion angle types with at least 10 hits per SMARTS. These statistics can be directly transformed into a torsion angle file for OMEGA.

Quantity-Quality Scan. To find good compromises for the two selection parameters described in the methods section—the minimum number of hits per SMARTS and the frequency cutoff for individual angle intervals, a systematic scan was performed between 10 and 100 hits/SMARTS in steps of 10 and between frequencies of 10% and 100% in 5% steps. This gave in total 170 experiments with different combinations of the two parameters. For each experiment, a torsion angle file was generated, and OMEGA was run over the 11 027 compounds in the test set. The results were analyzed and compared to the original X-ray structures. For each compound, the number of conformations and the RMS of the closest conformation compared with the crystal structure were recorded. For the entire test set of 11 027 compounds, the average number of conformations, the percentage of compounds with a conformation being 0.5 Å or closer to the X-ray structure, and the average RMS of the closest structure were calculated. Figure 2 shows a plot of the average number of conformations vs the average RMS value. Clearly, there is a compromise to be made between quantity and quality. Parameter sets that result in a higher number of conformations tend to result in lower RMS values. For finding a good compromise, we compared our results with results obtained using the OMEGA default torsion file (marked by a circle). The latter gave 50 conformations on average with an average RMS of 0.3 Å. When settling for about the same number of conformations and following the principle of Pareto optimality, the corresponding parameters for the CSD-based approach (min. 20 hits per SMARTS and a frequency cutoff of 45%) led to a substantially lower average RMS of 0.22 Å (marked by a circle in the plot). Setting the minimum number of hits per SMARTS to 20 reduced the number of SMARTS rules further from 2863 to a final number of 1864.

Table 1. Comparison of the Coverage of Conformer Generation by Corina and OMEGA with the Original Set of Torsion Rules and the New Torsion File from This Study

	Corina	Omega		Student
		default torlib	new torlib	*t*-test[b]
converted	11 016	8580[a]	8579[a]	
RMS < 0.5 Å	43%	84%	92%	
av RMS	0.74 Å	0.30 Å	0.22 Å	17.0
av conf	1	50	50	
CPU time	149 s	351 s	627 s	

[a] Structures with at least one new conformation generated by OMEGA. [b] Student *t*-test performed on the average RMS values obtained with the default and the new torlib rules.

X-ray Coverage. To assess the quality of the generated conformer ensembles, the coverage of the torsional space was measured by estimating the portion of structures in the CSD training set (11 027 compounds) which had at least one OMEGA conformation closer than 0.5 Å compared with the X-ray structure. This is a rough measure of coverage since on average the crystal conformations can be considered as a reasonable low-energy conformation. Table 1 shows the results of OMEGA runs with the default torsion file and the new set of torsion rules: the number of converted structures (for OMEGA, the number of structures with at least one new conformation), the percentage of structures with at least one conformation with an RMS < 0.5 Å, the average RMS of the closest OMEGA conformations, the average number of conformations, and the CPU time spent on a 2.4 GHz Pentium III computer. In addition, we show for comparison reasons the same data for the 3D-structure generator Corina[11] generating one conformation per molecule. As can clearly be seen, the multiple conformer ensembles generated by OMEGA improve significantly coverage and average RMS compared with the single-conformer set generated by Corina. We can also see a significant improvement of both criteria when running OMEGA with the new set of torsion rules. The number of structures with at least one conformation closer than 0.5 Å to the X-ray structure increases from 84% to 92% and the average RMS drops from 0.30 to 0.22 Å. The significance of the improvement is further supported by the high Student *t*-test value of 17.0 for the average RMS values. At the same time, the number of conformations remains at around 50. There is a price paid in terms of CPU time that increases from 351 to 637 s. This is due to the 15-fold increase of the number of SMARTS rules (123−1864).

Comparison between Test and Training Set. To assess the dependence of the coverage results from the training set we analyzed the results obtained with the new torsion rules for both the training set of 20 000 structures and the test set of 11 027 structures. Table 2 shows a comparison of the same criteria as in Table 1. Clearly, there is no significant

Table 2. Comparison of the Coverage of Conformer Generation by OMEGA with the New Torsion File for the Training Set and the Test Set

	training set	test set
converted[a]	15 599	8579
RMS < 0.5 Å	92%	92%
av RMS	0.22 Å	0.22 Å
av conf	49	50

[a] Structures with at least one new conformation generated by OMEGA.

Table 3. Analysis of the SMARTS Searching Performance for the Default Torsion File and the New One from This Study

	default torlib	new torlib
SMARTS	123	1864
not found	14 561 (1.3/mol)	1441 (0.13/mol)
SMARTS/mol	92.6	503
time (SMARTS search)	18 s	131 s

difference between the training and the test sets at all. Thus, the number and diversity of compounds in the training set seem to be sufficient in order to predict new data with the same quality.

Torsion Coverage. OMEGA has a set of unspecific general torsion rules for cases where no adequate SMARTS pattern can be applied. An important question to address is how much the improved torsion rules aid OMEGA to find more often a relevant SMARTS pattern. Table 3 shows for the test set of 11 027 X-ray structures the number of bonds where OMEGA reports a missing SMARTS pattern, the average number of SMARTS searched per molecule before an appropriate SMARTS was found (OMEGA stops at the first hit), and the time spent exclusively on the SMARTS searches. This was simulated by using an in-house program based on OELib assuming that the SMARTS procedures are quite similar compared with OMEGA. Clearly, the number of torsions without an appropriate SMARTS pattern decreases by an order of magnitude from 1.3 per molecule to 0.13 per molecule. This finding becomes even more relevant when comparing these figures to the average number of rotatable bonds per molecule in the test set. On average, OMEGA finds 2.6 rotatable bonds per molecule. Thus, using the default rules almost half of the torsions are not assigned a specific torsion rule! Of course, more SMARTS lead to longer search times—131 s instead of 18 s. This is due to the fact that OMEGA searches on average about 93 SMARTS per molecule with the default rules and more than 500 with the new SMARTS set.

Bioactive Conformations. Crystal structures as collected in the CSD are not necessarily representative for conformations that small molecules adopt when they bind to a biological target—the so-called bioactive conformation. To assess the coverage of conformational space when using bioactive conformations as reference, a set of 1267 ligands in their bound conformation in protein−ligand complexes from the PDB[12] was used. This is a subset of drug-sized ligands from PDB with corrected valence notations and protonation states generated for a previous study.[13] Table 4 summarizes the results. We used the same criteria as described above for the CSD data set selection. Clearly, this set behaves somewhat differently compared with the CSD set. The overall coverage is lower, whereas the average RMS as well as the average number of conformations is higher. This might simply be a result of a higher average size and flexibility of the molecules in this set (data not shown). However, the relevant finding is consistent with the results above—the new set of torsion rules improves coverage. The significance of this improvement is also supported by the rather high Student *t*-test value of 4.16 when comparing the average RMS values.

Table 4. Coverage of Conformer Generation for the Set of 1267 Ligands from the PDB

	default torlib	new torlib	Student *t*-test[b]
converted[a]	1167	1168	
RMS < 0.5 Å	70%	79%	
av RMS	0.40 Å	0.35 Å	4.16
av conf	80	94	

[a] Structures with at least one new conformation generated by OMEGA. [b] Student *t*-test performed on the average RMS values obtained with the default and the new torlib rules.

Table 5. Influence of MMFF Postoptimization of the Test Set

	default torlib		new torlib	
	no MMFF	MMFF	no MMFF	MMFF
RMS < 0.5 Å	84%	74%	92%	76%
av RMS	0.30 Å	0.39 Å	0.22 Å	0.37 Å

MMFF Influence. OMEGA comes with a built-in reimplementation of the MMFF force field[14] that can be used to postoptimize all generated conformations. A relevant question in this context was the potential improvement of the quality of the generated conformer ensembles (measured as above) and whether the use of MMFF might level out the differences between different torsion rules. To analyze this, OMEGA was run over the test set using both the default and the new set of SMARTS rules with the MMFF postoptimization switched on (“-finalopt true”). We tried also to combine this with a reduced MMFF interaction model without Coulomb and attractive van der Waals terms ("-truncmmff true”) in order to reduce intramolecular forces. The results were more or less the same (data not shown). Table 5 shows coverage and average RMS for the test set of 11 027 CSD structures, the two different torsion files (default and new), and MMFF postoptimization switched off and on. Clearly, the use of MMFF levels out the performance differences between the two sets of torsion rules. Unfortunately, this goes into the wrong direction with decreased quality. The percentage of structures with at least one OMEGA conformation closer than 0.5 Å compared with X-ray drops to 74−76%, and the average RMS increases to 0.37−0.39 Å. Thus, MMFF seems to favor a different part of conformational space than that occupied by X-ray structures. We found more or less the same tendency when postoptimizing the conformations generated for the PDB ligand set (data not shown).

CONCLUSIONS

We have presented an approach for the automatic derivation of torsion angle rules for the conformer generator OMEGA based on X-ray structures from the CSD. Our results show that automatically generated SMARTS for four-atomic torsional patterns can be used to search databases of experimental structures and to derive improved torsion angle rules for conformer generation. When compared to the default rules implemented in OMEGA, our new torsion data file leads to a significant improvement of coverage of torsional space from 84% to 92% and of the average RMS of the closest conformation from 0.30 to 0.22 Å for a test set of 11 027 crystal structures from the CSD. We did not see any improvement according to these criteria when postoptimizing the conformations with OMEGA's built-in MMFF force field.

The new set of torsion rules will in the future be used for our work when preparing multiconformer databases with OMEGA. Potentially, the new torsion rules can also be adapted to docking programs such as GOLD or FlexX that use similar torsion data files.

ACKNOWLEDGMENT

We acknowledge valuable comments and help from our colleagues Jeff Morris, Stefan Schmitt, and Dave Timms.

REFERENCES AND NOTES

(1) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(2) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein−Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609−623.

(3) (a) Sprague, P. W. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. *Perspect. Drug Discovery Des.* **1995**, *3*, 1−20. (b) Sprague, P. W.; Hoffman, R. Catalyst Pharmacophore Models and their Utility as Queries for Searching 3D Databases. In *Computer-Assisted Lead Finding and Optimization—Current Tools for Medicinal Chemistry;* van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; VHCA: Basel, 1990; pp 230−240.

(4) ROCS, FRED, OMEGA, OELib: OpenEye Scientific Software: Santa Fe, NM, 2005.

(5) Boström, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.

(6) Klebe, G.; Mietzner, T.; Weber, F. Methodological developments and strategies for a fast flexible superposition of drug-size molecules. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 35−49.

(7) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, J. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(8) *Daylight Software Manual*: *Theory*; Daylight Information Systems: Santa Fe, NM, 2005.

(9) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New software for searching the Cambridge Structural Database and visualising crystal structures. *Acta Crystallogr.* **2002**, *B58*, 389−397.

(10) SYBYL; Tripos Inc.: St. Louis, MO, U.S.A., 2005.

(11) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(12) Bernstein, F.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Schimanouchi, T.; Tasumi, M. J. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(13) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, 2004; pp 271−285.

(14) (a) Halgren, T. A. Merck Molecular Force Field I.-V. *J. Comput. Chem.* **1996**, *17*, 490−641. (b) Halgren, T. A. Merck Molecular Force Field VI.-VII. *J. Comput. Chem.* **1999**, *20*, 720−748.

CI060042S