Benchmark Data Set for in Silico Prediction of Ames Mutagenicity

Katja Hansen,[†] Sebastian Mika,[‡] Timon Schroeter,[†] Andreas Sutter,[§] Antonius ter Laak,[§] Thomas Steger-Hartmann,[§] Nikolaus Heinrich,[§] and Klaus-Robert Müller*,[†]

University of Technology, Berlin, Germany, idalab GmbH, Berlin, Germany, and Bayer Schering Pharma AG, Berlin, Germany

Received May 6, 2009

Up to now, publicly available data sets to build and evaluate Ames mutagenicity prediction tools have been very limited in terms of size and chemical space covered. In this report we describe a new unique public Ames mutagenicity data set comprising about 6500 nonconfidential compounds (available as SMILES strings and SDF) together with their biological activity. Three commercial tools (DEREK, MultiCASE, and an off-the-shelf Bayesian machine learner in Pipeline Pilot) are compared with four noncommercial machine learning implementations (Support Vector Machines, Random Forests, k-Nearest Neighbors, and Gaussian Processes) on the new benchmark data set.

1. INTRODUCTION

The bacterial reverse mutation assay (Ames test[1−4]) to detect mutagenicity in vitro is of crucial importance in drug discovery and development as an early alerting system for potential carcinogenicity and/or teratogenicity. Existing commercial tools suitable for predicting the outcome of the Ames test, such as DEREK[5] for Windows and MultiCASE,[6] provide promising results on several evaluation data sets[5,7−9] and the possibility to derive structure−activity and/or even mechanistic information from the predictions. Still, these two commercial tools are limited in terms of statistical performance, technical accessibility for bench chemists, and adaptability to a company's chemical space.

In the literature several approaches are described to predict Ames mutagenicity, generally yielding good specificity and sensitivity values (prediction accuracy of up to 85%). Depending on the descriptors and the statistical methods used, some of the models offer structure−activity information (such as Helma et al.[10] or Kazius et al.[11]). Some are however harder to interpret due to the choice of chemical descriptors derived from structural information (such as Feng et al.[12]). In order to identify an adaptable, well-performing and technically feasible prediction model for mutagenicity, a large and clearly defined benchmark data set is needed. We found that the underlying structural information of commercial prediction tools is at least in part not accessible and limited in terms of size and chemical space covered. Thus, the user cannot reproduce the deduction of structure−activity relationships (SARs) in all cases. Moreover, different data sets have been used in the scientific literature[10−12] without disclosing the splits (training set/test set) used for model evaluation. Hence, a reasonable comparison of different methods with these data appears impossible. To overcome this problem we collected a new benchmark set of 6512 compounds together with their Ames test results from public sources. As described in this report, we make this large unique benchmark set - including well-defined random splits - publicly available (see http://ml.cs.tu-berlin.de/toxbenchmark/) to facilitate future comparisons with prediction methods of other researchers.

2. THE AMES MUTAGENICITY BENCHMARK DATA SET

In the Ames test,[1−4] frame-shift mutations or base-pair substitutions may be detected by exposure of histidine-dependent strains of *Salmonella typhimurium* to a test compound. When these strains are exposed to a mutagen, reverse mutations that restore the functional capability of the bacteria to synthesize histidine enable bacterial colony growth on a medium deficient in histidine ("revertants"). Since many chemicals interact with genetic material only after metabolic activation by enzyme systems not available in the bacterial cell, the test compounds are in many cases additionally examined in the presence of a mammalian metabolizing system, which contains liver microsomes (with S9 mix).

A compound is classified Ames positive if it significantly induces revertant colony growth at least in one out of usually five strains, either in the presence or absence of S9 mix. A compound is judged negative if it does not induce significant revertant colony growth in any strain reported, both in the presence and absence of S9 mix. As a consequence of this definition, Ames negative compounds in the benchmark data set which have been tested in selected strains only may turn out to cause reverse mutations when being examined in additional strains. Overall, Ames positive compounds are thus more accurately defined in the benchmark data set as their label will not change with further testing.

The benchmark data set was derived from information contained in CCRIS,[13] Helma et al.,[10] Kazius et al.,[11] Feng et al.,[12] VITIC,[14] and GeneTox databases[15] using the Software PipelinePilot.[16] Duplicate structures have been removed. Additionally, a small number of 20 extraordinary and/or inorganic molecules have been omitted from the data

[*] Corresponding author e-mail: klaus-robert.mueller@tu-berlin.de.

[†] University of Technology.

[‡] idalab GmbH.

[§] Bayer Schering Pharma AG.

10.1021/ci900161g CCC: $40.75 © 2009 American Chemical Society
Published on Web 08/24/2009

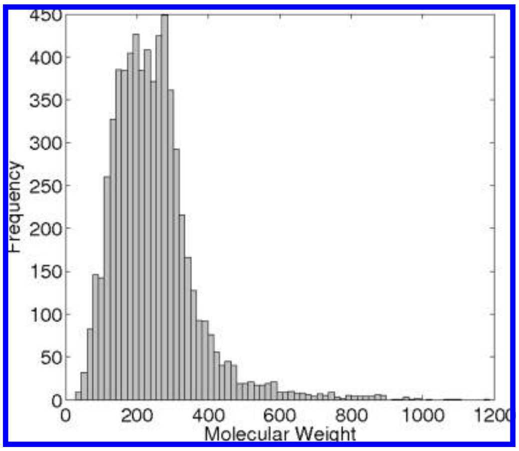


Figure 1. Distribution of molecular weights in the benchmark data set.

Table 1. Overview on the Number of Compounds and Results Taken from Each Source When Stepwise Extending the Ames Mutagenicity Data Set[a]

	AMES positives	AMES negatives	total
CCRIS[13]	1359	1180	2539
Kazius et al.[11]	1375	849	2224
Helma et al.[10]	81	57	138
Feng et al.[12]	280	111	391
VITIC[14]	386	808	1194
GeneTox[15]	22	4	26
total	3503	3009	6512

[a] Due to overlaps between different sources, the total number of relevant data contained in the individual databases can be higher.

set. Specified stereocenters, Chemical Abstracts Service (CAS) numbers, and World Drug Index (WDI) names are indicated, where possible. A small number of 25 compounds had contradictory experimental results with respect to DEREK or MultiCASE internal data and were thus removed from the data set. The final set contains 6512 compounds as canonical SMILES together with the corresponding Ames test results and references (see the Supporting Information and/or http://ml.cs.tu-berlin.de/toxbenchmark/). The benchmark data set includes 1414 compounds from the World Drug Index. The mean molecular weight of the data set is 248 ± 134 (Median MW: 229, see Figure 1).

An overview of the Mutagenicity Benchmark data set is presented in Table 1. The table gives the number of compounds and results taken from each source when stepwise extending the Ames mutagenicity data set. Due to overlaps between different sources, the total number of relevant data contained in the individual databases can be higher.

It is noted that these databases also contain data from Ames tests that were performed before strict regulatory requirements were imposed for the authorization of new chemicals. In general, the reproducibility of the Ames test is affected e.g. by the purity of the test compound, the bacterial strains and metabolic activation mixtures used, the experimental procedure, nonspecific effects such as cytotoxicity, and the interpretation of results.[2] Still, the experimental reproducibility of the Ames test was determined to be 85%.[17,18] Consequently, Ames data are generally affected by a 15% error. In agreement, when compiling their data set, Kazius et al. performed a consistency check between the original sources CCRIS and NTP, yielding an error rate of only 11% which is even below the average interlaboratory reproducibility error of Ames tests (the authors removed these contradictory results).

For the data taken from the papers by Kazius et al.,[11] Helma et al.,[10] and Feng et al.,[12] each individually compiled from different public sources, no information was available on test outcomes for the different strains with and without S9 mix. The classification rules were consistent between the papers and in accordance with international guidelines.[19] Consequently, for the data compiled from CCRIS, GENETOX, and VITIC which differentiate by strain and presence/absence of S9 mix, we applied the same standard classification rules. Thus, for the benchmark data set as a whole, additional information (experimental conditions, revertant frequencies in different strains, metabolic activation) is not available.

3. TOOLS AND METHODS

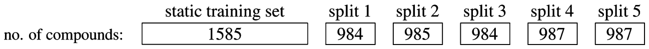
Using the benchmark data set, we evaluated four noncommercial implementations of machine learning techniques and three commercial prediction tools. For the noncommercial predictors compounds were represented as numerical vectors of molecular descriptors. Molecular descriptors were selected from blocks 1, 2, 6, 9, 12, 15, 16, 17, 18, and 20 of DragonX version 1.2[20] based on a 3D structure generated by Corina version 3.4.[21] These DragonX blocks provide a wide variety of descriptor types including constitutional, topological, geometrical, functional group count, and atom-centered fragments descriptors as well as various molecular properties, representing a starting point for unbiased modeling attempts on the new benchmark data set.

In the following we provide details on algorithms and implementations, i.e. for the noncommercial implementations of Support Vector Machines, Gaussian Processes, Random Forests, and k-Nearest Neighbors as well as for the commercial tools DEREK and MultiCase.

Support Vector Machines (SVMs) are one of the most popular methods in machine learning and have been applied successfully to classification tasks in computational chemistry.[22,23] The main advantage of SVMs is the sparsity of their solutions. Basically these methods project the input data into a high-dimensional feature space and then construct a hyper plane to separate the two classes with maximum margin.[24,25] Using kernel functions, complex nonlinear projections can be applied in this framework without significant additional computational efforts. Our implementation of this algorithm is based on libsvm[26] and uses the radial basis function kernel.

Gaussian Process classification is a technique from the field of Bayesian statistics which predicts the probability of a compound being mutagen instead of the classification itself.[27] Taking advantage of this concept a notion of the "domain of applicability" is provided in addition to the classification (see Schwaighofer et al.[28]). In contrast to the SVM and the Random Forest[29] the optimal model parameters can be inferred from the data without any external parameter selection method.

The **Random Forest** used in this approach is essentially a collection of 50 decision trees where each tree depends on a set of randomly selected features of the input data. The



	static training set	split 1	split 2	split 3	split 4	split 5
no. of compounds:	1585	984	985	984	987	987

Figure 2. Splitting of the benchmark data set for cross-validation.

decision tree itself recursively splits the data into subsets to separate the two classes. Our implementation is based on the work of Breimann et al.[29] but omits any bootstrapping or bagging of samples.

Additionally, a **k-Nearest Neighbor** model is implemented as a baseline model. For each new compound the k closest compounds within the training set are selected and the relative class frequency within this neighborhood forms the prediction.

Pipeline Pilot's Bayesian categorization model[16] provides supervised Bayesian learning for large data collections. In our example, this Bayesian classifier was combined with Pipeline Pilot's chemical fingerprint technology. Several modeling efforts showed that Extended Connectivity Fingerprints as developed by Sci-Tegic[16] (especially ECFP_4) yielded optimal Bayesian models for the current data set.

DEREK[5,30] is an expert system providing known structure−activity relationships (SARs), each relating to a toxicological end point. The mutagenicity prediction by DEREK was considered positive if a structure triggered at least one mutagenicity alert, and negative if it did not raise any mutagenicity alerts. Compounds contained as examples in the knowledge base of DEREK[5] were excluded from the evaluation due to technical considerations. *In practice, the DEREK example database contains knowledge essential for the evaluation of chemicals.*

MultiCASE[6,31] is a correlative tool predicting toxicity on the basis of structural fragments statistically correlated with activity (QSAR). For mutagenicity prediction, the commercially available AZ2 mutagenicity module was used. Compounds contained in the training set of the AZ2 module were excluded from the evaluation.

4. EVALUATION

Models were evaluated in a 5-fold cross-validation setting. First all compounds which were verifiable known to DEREK or MultiCASE were pooled together in a static training set (Figure 2). The remaining data set was divided into five cross-validation splits. Within each step of the cross-validation all models were trained on a set of four cross-validation splits together with the static training set (at least 5525 compounds). The fifth split forms the validation set. To select the parameters for the machine learning algorithms an inner loop of cross-validation was performed on the training sets.

We measured the quality of the resulting models using the Receiver Operating Characteristic[32] (ROC, see Figure 3). In an ROC graph the false positive rate (1 - specificity, see Table 2) is plotted against the true positive rate (sensitivity). The point (0,1) in this diagram marks a perfect classifier; at (0,0) all samples are classified as negative and in (1,1) all samples are assigned to the positive class.

For parametric classifiers ROC-Curves are drawn between the two latter points by varying the cutoff parameter. To compare the performance of parametric classifiers we use the area under this ROC-Curve (AUC). This measure can be interpreted as the probability that the classifier will assign a higher score to a positive example than to a negative one. Within the parameter selection process of the parametric classifiers the AUC is maximized. A nonparametric classifier results in a single point in the ROC graph. Here the graph illustrates the relation of sensitivity versus 1 - specificity.

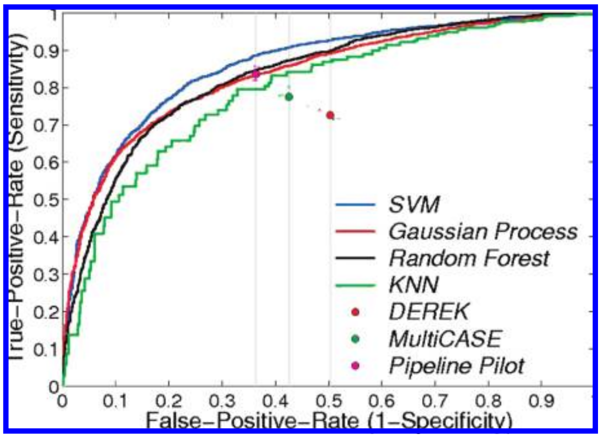


Figure 3. ROC curves for the different commercial and noncommercial models as estimated from the validation procedure described in the text.

Table 2. Relation of Performance Measures

		observation	
		positive	negative
prediction	positive	TP	FP
	negative	FN	TN
true positive rate = sensitivity = $(TP)/(TP + FN)$			
true negative rate = 1−specificity = $(TN)/(TN + FP)$			

Table 3. Cross-Validation Results for Parametric Classifiers (See Text for Details)

model	AUC
SVM	0.86 ± 0.01
GP	0.84 ± 0.01
Random Forest	0.83 ± 0.01
k-Nearest Neighbor	0.79 ± 0.01

5. RESULTS AND DISCUSSION

All cross-validation results are illustrated in Figure 3. Tables 3 and 4 summarize the results with respect to the different performance measures. PipelinePilot, trained with the developed data set, shows the best results of the three commercial tools followed by MultiCASE. The expert system DEREK yields the lowest sensitivity and specificity of all considered models (cf. Table 4). MultiCASE and DEREK cannot take advantage of the rich information provided by the training data. They are based on a fixed set of mainly 2D descriptors (MultiCASE) or a static system of rules derived from a largely unknown data set and expert knowledge (DEREK). Moreover, the rules contained in DEREK - although some taking e.g. inter-relationships between different structural features or physicochemical properties into account - may be too generic to reflect the influence of the chemical neighborhood of a functional group on their mutagenic activity. It can be assumed that there are unknown structure activity relationships which are not contained in the DEREK knowledge database. The employed

Table 4. Comparison of Parametric and Nonparametric Classifiers[a]

	specificity 0.50	specificity 0.57	specificity 0.64	model
sensitivity	0.93 ± 0.01	0.91 ± 0.01	0.88 ± 0.01	SVM
	0.89 ± 0.01	0.86 ± 0.01	0.83 ± 0.02	GP
	0.90 ± 0.02	0.87 ± 0.02	0.84 ± 0.03	Random Forest
	0.86 ± 0.02	0.86 ± 0.02	0.81 ± 0.02	k-Nearest Neighbor
	-	-	0.84 ± 0.02	PipelinePilot
	0.73 ± 0.01	-	-	DEREK
	-	0.78 ± 0.02	-	MultiCASE

[a] The table shows the sensitivity values for fixed levels of specificity. For parametric classifiers the sensitivity can be calculated for arbitrary levels of specificity. Non parametric classifiers yield only one level of sensitivity. The specificity values of 0.5, 0.57, and 0.64 correspond to false-positive-rates 0.5, 0.43, and 0.36. The tabled values of sensitivity correspond to the intersection points of the ROC curves and the vertical gray lines in Figure 3.

AZ2 model of MultiCASE cannot be adapted to a specific chemical space and therefore yields a lower prediction accuracy. Nevertheless, DEREK and MultiCASE are still essential for drug discovery and development as they provide structure−activity and/or mechanistic information essential for structure optimization and regulatory acceptance.

In contrast to the commercial tools, the machine learning algorithms exclusively derive their knowledge from the training data. The fact that none of the other tools could outperform one of the machine learning models (Figure 3) indicates the power of the latter approaches and the high information content of the provided benchmark data set. The rather good performance of the simple k-Nearest Neighbor model indicates a strong influence of small local molecular changes on Ames mutagenicity. However the application of more sophisticated machine learning methods results in a significant performance gain especially for the support vector machine.

6. CONCLUSION

All five evaluated machine learning methods (SVM, Gaussian Process, Random Forest, k-Nearest Neighbors, and the commercial Pipeline Pilot) yield good results on the benchmark data set.

The future evaluation of additional prediction methods on the proposed benchmark data set represents a promising strategy for further optimization of Ames mutagenicity prediction. In addition, scientists interested in method development may benefit from the present work as all modeling and evaluation results obtained using our data set allow for a direct comparison of different methods.

From an industry perspective, the benchmark data set helped us to establish an adaptable and technically practical prediction model that statistically outperforms commercial predictive tools like DEREK and MultiCASE, yet the application of these standard tools is still justified to obtain interpretable structural information on the mutagenicity prediction.

Future development will strive for improving the accuracy of machine learning based prediction tools that yield at the same time interpretable results.

ACKNOWLEDGMENT

The authors wish to thank Eugen Muratov and Alexander Tropsha for their helpful comments on the data set. This work was supported in part by the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886 and DFG grant MU 987/4-1.

Supporting Information Available: Compounds as canonical SMILES together with the corresponding Ames test results and references. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ames, B. N.; Lee, F. D.; Durston, W. E. An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 782–786.

(2) Mortelmans, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **2000**, *455*, 29–60. PMID: 11113466.

(3) McCann, J.; Ames, B. N. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals: discussion. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73*, 950–954.

(4) McCann, J.; Spingarn, N. E.; Kobori, J.; Ames, B. N. Detection of carcinogens as mutagens: bacterial tester strains with R factor plasmids. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 979–983.

(5) Sanderson, D.; Earnshaw, C. Computer prediction of possible toxic action from chemical structure, the DEREK system. *Hum. Exp. Toxicol.* **1991**, *10*, 261–273.

(6) Klopman, G. MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176–184.

(7) Snyder, R. D. An update on the genotoxicity and carcinogenicity of marketed pharmaceuticals with reference to in silico predictivity. *Environ. Mol. Mutagen.* **2009**, *50*, 435–450, PMID: 19334052.

(8) Benfenati, E.; Gini, G. Computational predictive programs (expert systems) in toxicology. *Toxicology* **1997**, *119*, 213–225.

(9) Zeiger, E.; Ashby, J.; Bakale, G.; Enslein, K.; Klopman, G.; Rosenkranz, H. Prediction of Salmonella mutagenicity. *Mutagenesis* **1996**, *11*, 471–484.

(10) Helma, C.; Cramer, T.; Kramer, S.; Raedt, L. D. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402–11, PMID:15272848.

(11) Kazius, J.; Mcguire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312–320.

(12) Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y.; Yuan, S.; Young, S. S. Predictive toxicology: benchmarking molecular descriptors and statistical methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1463–70.

(13) *Chemical Carcinogenesis Research Information System*; 2009; on the NCRI Informatics Initiative Homepage. http://www.cancerinformatics.org.uk/matrix/CCRIS.htm (accessed July 17, 2009).

(14) Judson, P. N.; et al. Towards the creation of an international toxicology information centre. *Toxicology* **2005**, *213*, 117–28. PMID: 16084005.

(15) *Genetic Toxicity, Reproductive and Developmental Toxicity, and Carcinogenicity Database*; 2009, at http://www.fda.gov/AboutFDA/CentersOffices/CDER/ucm092217.htm (accessed July 17, 2009).

(16) Inc., A. S. *SciTegic Pipeline Pilot, version Version 7.0*; 2009; available at http://accelrys.com/products/scitegic/ (accessed August 17, 2009).

(17) Piegorsch, W. W.; Zeiger, E. Measuring intra-assay agreement for the Ames Salmonella assay. *Lecture Notes in Medical Informatics* **1991**, *43*, 35–41. Statistical Methods in Toxicology, Hothorn, L., Ed.; Lecture Notes in Medical Informatics, Springer-Verlag: Heidelberg, 1991; Vol. 43, pp 35−41.

(18) Benigni, R.; Giuliani, A. Computer-assisted analysis of interlaboratory Ames test variability. *J. Toxicol. Environ. Health* **1988**, *25*, 135–148.

(19) *OECD guidline for testing of chemicals: Bacterial Reverse Mutation Test*; 2009; on Web. http://www.oecd.org/dataoecd/18/31/1948418.pdf (accessed Jul 17, 2009).

(20) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, 1st ed.; Wiley-VCH: 2002.

(21) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(22) Ivanciuc, O. In *Applications of Support Vector Machines in Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley-VCH: 2007; Vol. 23, Chapter 6, pp 291−400.

(23) Müller, K.-R.; Rätsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'Drug-likeness' with Kernel-Based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–253.

(24) Müller, K.-R.; Mika, S.; Rätsch, G.; Tsuda, K.; Schölkopf, B. An Introduction to Kernelbased Learning Algorithms. *IEEE Neural Networks* **2001**, *12*, 181–201.

(25) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press: Cambridge, 2002.

(26) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; 2001; available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (accessed Jul 17, 2009).

(27) Rasmussen, C. E.; Williams, C. K. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, 2005.

(28) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; ter Laak, A.; Sülzle, D.; Ganzer, U.; Heinrich, N.; Müller, K.-R. Accurate Solubility Prediction with Error Bars for Electrolytes: A Machine Learning Approach. *J. Chem. Inf. Model.* **2007**, *47*, 407–424.

(29) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(30) Lhasa Ltd. *DEREK for Windows, Version 10.0.2 Service Pack 3, Knowledge Base Release DfW 10.0.0_25_07_2007*; 2007; Leeds, U.K.

(31) Multicase Inc. *MultiCASE, version 2.1*; Beachwood, OH, U.S.A., 2007.

(32) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.

CI900161G