Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 15 (2005) 2932–2937
Bioorganic & Medicinal Chemistry Letters

2D RNA-QSAR: assigning ACC oxidase family membership with stochastic molecular descriptors; isolation and prediction of a sequence from *Psidium guajava* L

Humberto González-Díaz,[a,b,*] Guillermín Agüero-Chapin,[b] Javier Varona-Santos,[c] Reinaldo Molina,[b,d] Gustavo de la Riva[e] and Eugenio Uriarte[a,*]

[a]*Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain*
[b]*CBQ and CAP, Faculty of Chemistry and Pharmacy, Central University of 'Las Villas' 54830, Cuba*
[c]*Biomedicine Unit, FES Iztacala, UNAM, Los Barrios Avenue Num1, Los Reyes Iztacala, Tlalnepantla DF 54090, Mexico*
[d]*Universität Rostock, FB Chemie, Albert-Einstein-Straße 3a, D 18059 Rostock, Germany*
[e]*Department of Molecular Microbiology, Institute of Biotechnology, UNAM, Apartado Postal 510-3, Cuernavaca 62250, Morelos, Mexico*

Received 13 December 2004; revised 3 March 2005; accepted 4 March 2005
Available online 5 May 2005

Abstract—Quantitative structure–activity relationship (QSAR) techniques for small molecules could be applied to nucleic acids. Unfortunately, almost all molecular descriptors are more successful at encoding branching information than sequences and/or cannot be back-projected. A solution for scaling the QSAR problem up to RNA may be to transform sequences into secondary structures first. Our group has used Markovian negentropies as molecular descriptors for drug design with preliminary results in bioinformatics [*Bioinformatics* **2003**, *19*, 2079]. However, RNA-QSAR studies on RNA molecules have not been described to date. Novel Markovian negentropies have been introduced here as molecular descriptors for 2D-RNA structures. An RNA-QSAR study of the ACC proteins from different plants has been carried out. The QSAR recognizes 19/20 sequences (95.0%) within the ACC family and 12/17 (70.6%) of the control group sequences. The model has a high Matthews' regression coefficient ($C = 0.68$). Overall cross-validation average accuracies were 14 out of 15 for ACC sequences (93.3%) and 10 out of 13 for control sequences (76.9%). Finally, ACC oxidase family membership was assigned to a new sequence isolated for the first time in this work from *Psidium guajava* L. A backprojection map for this sequence identifies the left stem (40%) and the main stem (45%) as highly important substructures. Results of an nBLAST experiment are consistent with this finding and indicate a high conservation score (>70) for left stem and main stem; whereas major loop, right stem, cap and major loop right half were hardly conserved.

1. Introduction

Studies in molecular biology have been generating more and more structural information that is ideal for study by cheminformatics techniques.[1] In particular, the bioorganic chemistry of RNA has generated a large amount of information in recent years.[2,3] As a consequence, different mathematical methods have been combined with computational approaches in genome analysis. Very interesting results have recently been reported by Grau and co-workers concerning genome algebra.[4,5] In this sense, Markov models (MM) represent another well-known tool for the analysis of biological sequence data and these approaches have been used to study new genes and proteins.[6–8] However, in spite of its great potential to generate macromolecular structural descriptors for bioorganic chemistry, research MM have never been used to generate RNA molecular descriptors.

The use of molecular descriptors to derive quantitative structure–activity relationships (QSAR) is an approach of major interest. Molecular descriptors are numerical indices that codify either molecular or macromolecular structure. For instance, very interesting results have been reported by Gónzalez and co-workers on the

Keywords: RNA; QSAR; Markov model; Entropy; *Psidium guajava* L; Stochastic molecular descriptors.
* Corresponding authors. Tel.: +34 981563100x14938; fax: +34 981594912; e-mail addresses: humbertogd@vodafone.es; qofuri@usc.es

0960-894X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2005.03.017

application of molecular descriptors in bioorganic chemistry and biopolymer sciences.[9–15] Several specific and very successful indices (for small molecules) that use the concept of Shannon's entropy from the point of view of information theory have proven to be very effective in drug design.[16,17] For example, in the 1980's Kier used the concept of entropy to codify molecular structure through the so-called Molecular Negentropy in QSAR studies.[18]

Nevertheless, the combination of MM and entropy-like molecular descriptors in order to scale-up the classic QSAR problem from small-to-medium sized molecules toward a sort of RNA-QSAR have barely been studied. Very few RNA-QSAR studies have been reported, despite the real tendency to use analogues of molecular descriptors for small-to-medium sized in the evaluation of DNA and protein sequences.[19] Unfortunately, almost all molecular descriptors are more successful at encoding branching information and/or can not be back projected. In fact, the application of classical QSAR studies in bioorganic and medicinal chemistry often involves numerous branched molecules rather than linear ones. However, these approaches commonly use back-projectable descriptors as such as those in studies described by Roy et al., Cabrera-Pérez and co-workers.[20–22] Back-projection is the possibility of drawing a map that depicts the influence of every molecular substructure on the property under investigation.[23] A convenient solution to this problem could be to transform linear sequences into branched representations prior to the calculation of RNA back-projectable molecular descriptors. In this context, it would be useful to transform linear sequences into more branched representations prior to calculating the molecular descriptors, such as in the 2D representation investigated by Mathews et al.[24] This transformation of a 1D sequence into the 2D branched representation enables the encoding of more useful information with molecular descriptors and these data can be used to elucidate the QSAR.

Our group has largely used different stochastic molecular descriptors derived with Markov models to describe various biological activities of drugs, proteins and nucleic acids in the field of bioorganic medicinal chemistry.[25–28] In particular, entropy molecular descriptors called Markovian negentropies have been used by us as molecular descriptors for drug design and have provided interesting results in proteomics and bioorganic chemistry.[29–31] A preliminary report on RNA activity has also been published. This first study on bioinformatics focused on the local properties of a single RNA molecule.[32] However, QSAR studies concerning the global properties of several RNA molecules have not been reported to date.

The motivation for the present work stems from the use in QSAR of Markovian negentropies as molecular descriptors derived from RNA secondary structure. This process involves the introduction of new molecular negentropies for RNA secondary structure and these data are used to assign 1-aminocyclopropane-1-carboxylate (ACC) oxidase and synthase family membership,[33] which constitutes a family of proteins that has never previously been studied using RNA-QSAR techniques. For this reason we selected these specific properties for the RNA-QSAR study.

2. Results

The molecular biology of fruit maturation and ripening is currently of major importance in biotechnology.[34] In this processes ethylene regulation plays a significant role related to the ACC synthase and oxidase protein families.[35] In the first instance, the statistical significance of the results must be discussed before drawing any conclusions concerning the biology involved. Linear discriminant analysis (LDA)[36] was used to classify RNA sequences as ACC oxidase family members or not. In the development of the LDA the output was a dummy variable (ACCactv), which codifies whether a sequence lies within the ACC oxidase family (ACCactv = 1) or belongs to the control group (ACCactv = 0). In this problem the input was Θ_k, with k in the range [0, 5]. The software STATISTICA 6.0[37] was used and the best equation found to discriminate between ACC oxidase family members and the control group was:

$$\begin{aligned} &\text{ACCactv} = 2.88 \cdot {}^{1}O(\Theta_1) + 2.11 \cdot {}^{2}O(\Theta_3) - 9.04 \\ &N = 37 \quad C = 0.68 \quad F = 11.69 \quad p < 0.001 \end{aligned} \tag{1}$$

where λ is Wilk's statistic, N is the number of RNA sequences studied, F is Fisher's statistic and p is the p-level (probability of error). The symbols ${}^{I}O(\Theta_k)$ were used and, in this case, the superscript I expresses the order of importance of the variable (Θ_k) after a preliminary forward stepwise analysis and O signifies orthogonal. The p-level of Fisher's test for the LDA was <0.05. This means that the hypothesis of groups overlapping with a 5% error can be rejected.[36] This linear discriminant function classified correctly 81.1% of ACC oxidase/control group RNA sequences. More specifically, the model recognizes 19/20 sequences (95.0%) within the ACC oxidase family and 12/17 (70.6%) of the control group sequences. The high Matthews' regression coefficient ($C = 0.68$) points to a strong linear relation between the molecular descriptors and the ACC oxidase family membership classifications.[38] The coefficient for the LDA model $\rho = N/(\text{Nv} + 1) \times \text{Ng} = 37/(2 + 1) \times 2 = 6.17$ where the number of variables was Nv = 2 and the number of groups Ng = 2. This coefficient controls model over-fitting by taking into consideration the ratio of the number of sequences (N) with respect to the adjustable parameters (Nv and Ng), which has to be >4.[39] The step-by-step results for the forward stepwise analysis used to select the variables for the LDA are depicted in Table 1A in the supplementary data file. A high degree of collinearity was detected between Θ_k and the dependent. As a result, the variables were orthogonalized according to Randić's orthogonalization procedure,[40] mean-centred (with the arithmetic mean) and scaled with $1/\text{sd}(x_i)^2$, the inverse of the squared standard deviation.[41]

Jack-knife cross-validation (CV) experiments were performed by leaving out four different groups, which were

selected at random and contained 25% of the sequences. All of the statistical parameters for the new models—obtained after removing the RNA sequences—were checked in order to assess the stability of the model. Overall accuracies for those models and the average cross-validation accuracy (CV-average) were as follows: CV1 = 89.3%, CV2 = 88.9%, CV3 = 85.7%, CV4 = 85.7%, with an average CV value of 85.7. Details of the classification matrices and other parameters for training and cross validation experiments are reported in Table 2A of the supplementary material data file. In addition, more detailed results, including the names of the sequences used and their subsequent probabilities, can be found in Table 3A in the supplementary material.

As explained above, Θ_k, as defined here, required the use of the Shannon concept of entropy as described by Richard and Kier (1980).[17,18] The present model shows that the higher the entropy of interaction between bound nucleotides (covalent and hydrogen bonds) inside the secondary RNA backbone,[42] the higher the probability that a sequence is an ACC oxidase family member (by a factor of 2.88). Interestingly, the model predicts a similar influence (2.11) for indirect interactions between nucleotides that are not bound but placed at a topological distance equal to 3. An influence for long-range electrostatic interactions at topological distances >3 was not observed. Therefore, it is short-range electrostatic interactions rather than long-range ones that seem to control the likelihood that an RNA sequence will encode a protein of the ACC oxidase family. These results justify the use of a truncation function δ_{ij} in the present model in order to simplify calculations in place of a continuous treatment for the electrostatic field (see Models section).[43] In general, these results are consistent with those reported by Ramos de armos et al., who used Markovian negentropies in a study of the theoretical bioorganic chemistry of peptides and proteins.[44,45]

On the other hand, one of the must notable advantages of the present QSAR approach is the interpretation of the results in terms of the influence that each substructure has on the property in question. This objective is feasible by applying so-called back-projection analysis. This approach allows the QSAR model to be projected back onto the secondary RNA structure and is described in the method section.[46] In this work, the back-projection analysis of the RNA-QSAR (Eq. 1) is exemplified with a new sequence isolated from *Psidium guajava* L. This new RNA sequence was isolated by following well-established experimental protocols and has been published online by Agüero-Chapin et al. in the GENBANK but has not previously been published in a printed journal.[47] The results for total RNA isolated from dwarf guava (a) and RT-PCR reactions (b) are shown in Figure 1.

Our model was applied to the prediction of the average probability with which this sequence codifies a protein of the ACC oxidase family and this illustrates the use of the model ($P = 83.7\%$). Moreover, the back-projection study of the sequence was also carried out. Initially, the secondary structure was predicted with the RNASTRUCTURE 4.0 software developed by Mathews et al.[48] BIOMARKS 1.0® (*BIO*informatics *MARK*ovian *S*tudio) software, which has recently been developed in our laboratory,[49] was subsequently used to calculate the total Θ_k values for the sequence as a whole as well as for some specific substructures. BIOMARKS 1.0® uses as inputs the ct files generated by RNASTRUCTURE 4.0 and these files contain information on the RNA secondary structure. Finally, direct substitution into the RNA-QSAR model provides the contributions of each substructure, which were scaled into the range 0–100.

The back-projection results for this newly isolated sequence are represented in Figure 2 and are of particular interest due to the biotechnological importance of the maturation process of tropical fruits.[50] The results of the back-projection analysis need to be confirmed by direct experience and/or comparison with other theoretical approaches. A few back-projectable QSAR models

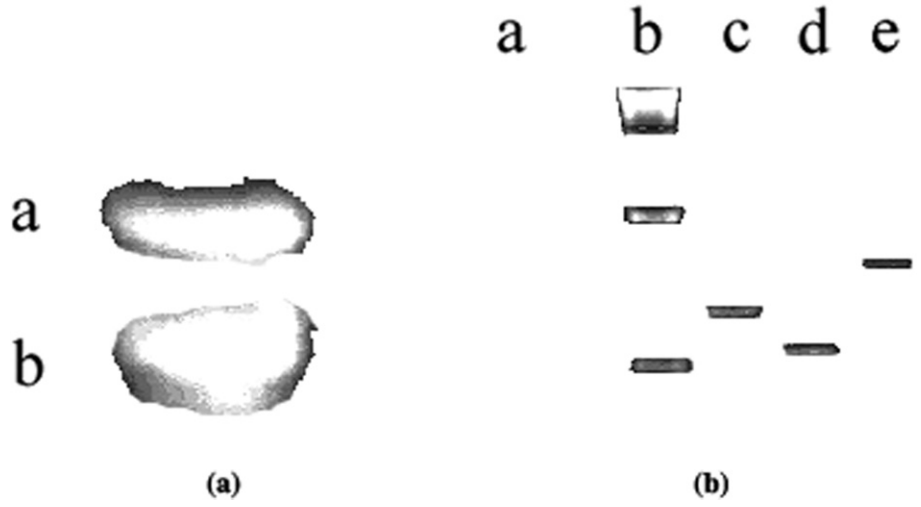

Figure 1. (a) Total RNA isolated from dwarf guava fruit mesocarps; (b) [a]negative control, [b]1Kb ladder (Gibco BLR), [c]RT-PCR reaction with degenerated primers, [d]positive control for Ready To Go™ RT-PCR Beads system, [e]PCR reaction with degenerated primers using genomic DNA isolated from the guava plant.

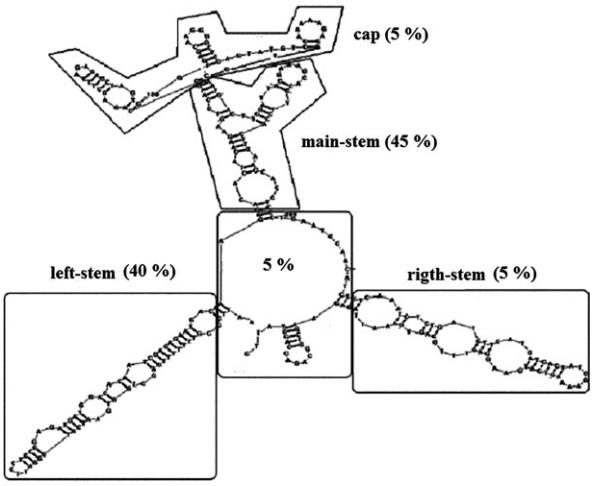

Figure 2. Backprojection map for guava ACC oxidase RNA secondary structure.

reported by Stief and Baunmann,[23] Cabrera-Pérez et al.,[51] and our group[52] have been applied to solve bioorganic chemistry problems and these stand out because of their good validation level.

In the present back-projection map analysis, the major importance in terms of the biological activity is predicted for the left stem (40%) and the main stem (45%). Conversely, other sub-structural patterns, such as the major loop and the right stem of this RNA sequence, are predicted to be less important. These results were confirmed by an nBLAST experiment, which is also reported here.[53] The nBLAST experiment showed a high conservation score (>70) for the left stem and main stem; whereas major loop, right stem, cap and major loop right half were hardly conserved. Further experiments will be carried out in subsequent, more in-depth studies in an effort to corroborate the prediction reported here (Fig. 3). However, given that the nBLAST technique is generally a successful method for other RNA sequences, it is promising that these results are in agreement with the model reported here. (see Fig. 1A, Supplementary data).

Numerous researchers throughout the world have applied QSAR techniques to solve bioorganic problems; see, for example, the work of Roy and Leonard.[54] Other previous studies have attempted to address the RNA structure–activity problem with molecular descriptors, but these were limited to a local situation for a single RNA molecule. To the best of our knowledge, the first two studies of this type were reported by González-Díaz and co-workers.[32,55] More recently, Marrero-Ponce et al. reported a study concerning the same problem.[56] The work described here opens up a new way to apply QSAR studies to several molecules and addresses a classical field of research, namely the RNA-QSAR problem. In particular, this study demonstrates the high versatility of the stochastic molecular descriptors developed by our group for bioorganic and medicinal chemistry studies.[57,58]

3. Models

In the work described here, the MARCH-INSIDE methodology was generalized to encode the 2D-RNA structure taking into account long-range electrostatic interactions. Direct long-range interactions through space are forbidden or the same truncated[59] long-range interactions are allowed to propagate step-by-step throughout the 2D-RNA ribbon. This approach used an MM to describe this propagation of long-range electrostatic interactions in a discrete step-by-step manner. Accordingly, the nucleotide–nucleotide short-range

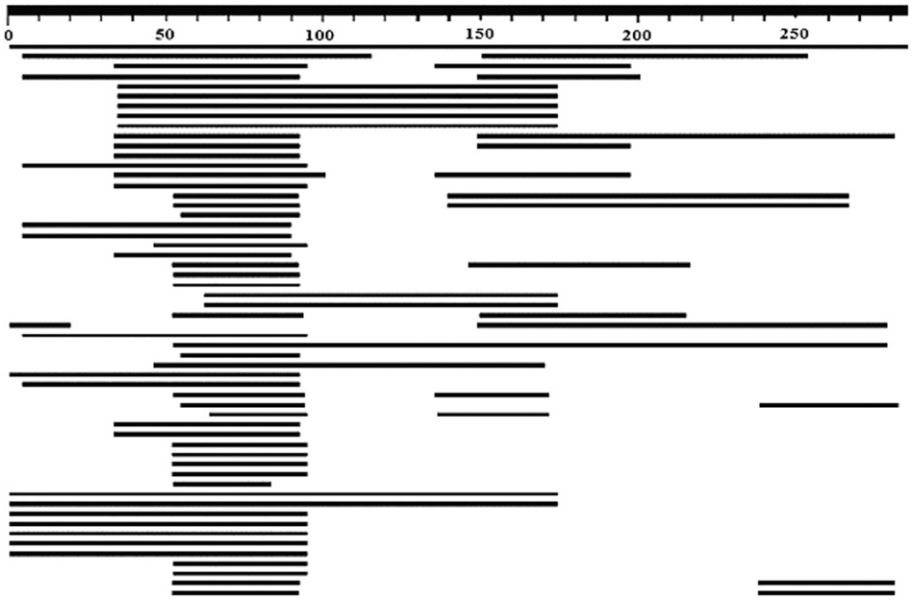

Figure 3. Results for the nBLAST experiment; the top rule scales number of bp, black bars indicate conserved regions for different sequences. Sequence names are not depicted.

electrostatic interaction ${}^{1}\Pi$ matrix (with elements p_{ij}) was used. ${}^{1}\Pi$ was built as a squared table of order n, where n represents the number of nucleotides in the RNA molecule.[32,55]

The elements of ${}^{1}\Pi$ (${}^{1}p_{ij}$), defined to codify information about the electrostatic interactions between nucleotides, were defined as:

$$ {}^{1}p_{ij} = \frac{\delta_{ij} \cdot \frac{q_j}{d_{ij}}}{\sum_{k=1}^{\alpha+1} \delta_{ij} \cdot \frac{q_j}{d_{ij}}} \quad (2) $$

where, δ_{ij} is the Kronecker symbol, which equals 1 for covalently or hydrogen bonded nucleotides and 0 otherwise; q_j is the electrostatic charge for the jth nucleotide; and d_{ij} is the topological distance, which is always equal to 1 due to δ_{ij} cutting-off of long-range interactions. The sum is carried out over all α nucleotides that interact directly—see the previous publication in this series for details. Once the stochastic matrix is defined for short-range electrostatic interactions, the Chapman–Kolgomorov equations were used to calculate the vector ${}^{A}\Pi_k$ of absolute probabilities ${}^{A}p_k(j)$ with which these short range electrostatic interactions propagate in a step-by-step manner and reach every jth nucleotide at distance k within the 2D-RNA framework, thus resulting in long-range indirect interactions between nucleotides (see references for similar models):[25–32]

$$ {}^{A}\Pi_k = {}^{A}\Pi_0 \times ({}^{1}\Pi)^k \quad (3) $$

where ${}^{A}\Pi_k$ is the vector of the initial probabilities ${}^{A}p_0(j)$ with which the jth nucleotide begins an interaction and can be calculated using Eq. 1 but summing up the n nucleotides instead of α. As ${}^{A}p_k(j)$ depends on the specific nucleotides (identified by q_j) and on the connectivity between the nucleotides in the RNA molecule, we can assert that any function having ${}^{A}p_k(j)$ values as arguments can encode information on 2D-RNA structure.[44,45,52] Functions classically used to encode information in QSAR are the Shannon entropy functions such as molecular negentropies. In this sense, we introduce here for the first time the average Electrostatic Markovian Molecular Negentropies as 2D-RNA backbone molecular descriptors (Θ_k) (note the analogies with our previous negentropies for small molecules).[60] The calculation of Θ_k was carried out using our experimental software BIOMARKS 1.0®. These parameters represent the entropy of electrostatic interaction for nucleotides at a topological distance equal to k or less:

$$ \Theta_k = -\sum_{j=1}^{n} {}^{A}p_k(j) \log \left({}^{A}p_k(j) \right) \quad (4) $$

4. Generation of descriptors and software implementation status

The calculation of Θ_k for short-to-middle length RNA secondary structures (having 2 to around 600 bp) was implemented in an updated version of our experimental software BIOMARKS 1.0® (*BIO*informatics *MARK*ovian *S*tudies).[49] This software has a connectivity table input modulus, which uploads the ct files generated with the RNASTRUCTURE 4.0 software developed by Mathews et al.[48] It is subsequently possible to select the calculation option and perform the calculation of molecular indices. RNA sequences that are significantly larger than 600 bp cannot be handled at present.

5. Experimental

All the steps for isolation and characterization of the new sequences isolated from *Psidium guajava* L were carried out according to well-established experimental protocols, which include: (1) collection of vegetable matter, (2) nucleic acid extraction, (3) primer design, (4) RT-PCR, (5) PCR reaction, (6) cloning and (7) sequencing, see Supplementary data for details.[61,62]

Acknowledgements

H. González-Díaz expresses his gratitude to the Laboratory of Medicinal Chemistry, Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, Spain, for financial support.

Supplementary data

Supplementary data associated with this article can be found, in the online version at doi:10.1016/j.bmcl.2005.03.017.

References and notes

1. Vázquez-Padron, R. I.; de la Riva, G.; Agüero, G.; Silva, Y.; Pham, S. M.; Soberón, M.; Bravo, A.; Aïtouche, A. *FEBS Lett.* **2004**, *578*, 30.
2. McPike, P. M.; Goodisman, J.; Dabrowiak, C. J. *Bioorg. Med. Chem.* **2002**, *10*, 3663.
3. Sullivan, J. M.; Goodisman, J.; Dabrowiak, C. J. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 615.
4. Sánchez, R.; Morgado, E.; Grau, R. *WSEAS Trans. Biol. Biomed.* **2004**, *1*, 190.
5. Sánchez, R.; Morgado, E.; Grau, R. *MATCH* **2004**, *52*, 29.
6. Chou, K. C. *Biopolymers* **1997**, *42*, 837.
7. Borodovsky, M.; Macininch, J. D.; Koonin, E. V.; Rudd, K. E.; Médigue, C.; Danchin, A. *Nucleic Acid Res.* **1995**, *23*, 3554.
8. Yuan, Z. *FEBS Lett.* **1999**, *573*, 23.
9. González, M. P.; Moldes, M. C. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 3077.
10. González, M. P.; Moldes, M. C. *Bioorg. Med. Chem.* **2004**, *12*, 2985.
11. González, M. P.; Días, L. C.; Morales, A. H.; Rodríguez, Y. M.; Gonzaga de Oliveira, L.; Gómez, L. T.; González-Díaz, H. *Bioorg. Med. Chem.* **2004**, *12*, 4467.
12. González, M. P.; Días, L. C.; Morales, A. H. *Polymer* **2004**, *45*, 5353.
13. González, M. P.; Moldes, M. C. *Bull. Math. Biol.* **2004**, *66*, 907.

14. González, M. P.; Morales, A. H.; Molina, R.; García, J. F. *Polymer* **2004**, *45*, 2773.
15. Morales, A. H.; González, M. P.; Rieumont, J. B. *Polymer* **2004**, *45*, 2045.
16. Bonchev, D.; Trinjastic, N. *J. Chem.* **1977**, *67*, 4517.
17. Shannon, C. E. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1955.
18. Kier, L. B. *J. Pharm. Sci.* **1980**, *69*, 807.
19. Randić, M.; Vračko, M.; Nandy, A.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235.
20. Roy, K.; Chakraborty, S.; Saha, A. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3753.
21. Cabrera-Pérez, M. A.; Bermejo, M.; González, M. P.; Ramos, R. *J. Pharm. Sci.* **2004**, *7*, 1701.
22. Cabrera-Pérez, M. A.; García, A. R.; Teruel, C. F.; Alvarez, I. G.; Bermejo-Sanz, M. *Eur. J. Pharm. Biopharm.* **2003**, *56*, 197.
23. Stief, N.; Baumann, K. *J. Med. Chem.* **2003**, *46*, 1390.
24. Mathews, D. H.; Turner, D. H.; Zuker, M. RNA Secondary Structure Prediction. In *Current Protocols in Nucleic Acid Chemistry*; Beaucage, S., Bergstrom, D. E., Glick, G. D., Jones, R. A., Eds.; John Wiley and Sons: NY, 2000.
25. González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. *Bioorg. Med. Chem.* **2004**, *13*, 323.
26. González-Díaz, H.; Olazábal, E.; Castañedo, N.; Hernández, S. I.; Morales, A.; Serrano, H. S.; González, J.; Ramos de Armas, R. *J. Mol. Model.* **2002**, *8*, 237.
27. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos de Armas, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model.* **2003**, *9*, 395.
28. González-Díaz, H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
29. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de Armas, R. *Bull. Math. Biol.* **2004**, *66*, 1285.
30. González-Díaz, H.; Molina, R.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691.
31. González-Díaz, H.; Molina, R.; Uriarte, E. *Polymer* **2004**, *45*, 3845.
32. González-Díaz, H.; Ramos de Armas, R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
33. Castellano, J. M.; Vioque, B. *Plant Growth Reg.* **2002**, *38*, 203.
34. Giovannoni, J. *Ann. Rev. Plant Phys. Plant Mol. Biol.* **2001**, *52*, 725.
35. Yueming, J. *Plant Growth Reg.* **2000**, *30*, 193.
36. Manhnhold, R., Krogsgaard, L., Timmerman, H., Eds. Chemometric Methods in Molecular Design; Van Waterbeemd, H., Ed.; VCH: Weinheim, 1995; Vol. 2.
37. Statsoft Inc. STATISTICA for windows, version 6.0; 2002.
38. Matthews, B. W. *Biochim. Biophys. Acta* **1975**, *405*, 442.
39. García-Domenech, R.; de Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445.
40. Randić, M. *J. Comput. Chem.* **1993**, *14*, 363.
41. Geladi, P.; Kowalski, B. R. *Anal. Chem. Acta* **1986**, *185*, 1.
42. Mathews, D. H.; Zuker, M. RNA Secondary Structure Prediction. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*; Clote, P., Ed.; John Wiley and Sons: NY, 2004.
43. Norberg, J.; Nilsson, L. *Acc. Chem. Res.* **2002**, *35*, 465.
44. Ramos de Armas, R.; González-Díaz, H.; Molina, R.; González, M. P.; Uriarte, E. *Bioorg. Med. Chem.* **2004**, *12*, 4815.
45. Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E. *Protein: Struct. Funct. Bioinf.* **2004**, *56*, 715.
46. Mathews, D. H.; Zuker, M. Predictive Methods Using RNA Sequences. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*; Baxevanis, A., Ouellette, F., Eds.; John Wiley and Sons: NY, 2003.
47. Agüero-Chapin, G.; Rodriguez, E. A.; Varona-Santos, J.; Díaz-Asencio, L.; Bacallao-Díaz, N.; Jiménez-González, E. Psidium Guajava Fruit Ripening-Related ACC oxidase mRNA, Partial cds., *GenBank*, **2002**, *Accession number*: AY123201. http://www.ncbi.nlm.nih.gov.
48. Mathews, D. H.; Zuker, M.; Turner, D. H. RNAStructure©, version 4.0; 2002.
49. González-Díaz, H.; Molina, R.; Sánchez, I. BIOMARKS 1.0©(BIOinformatics MARKovian Studies), version 1; 2004. The current version is experimental; please e-mail corresponding authors for details: humbertogd@vodafone.es or qofuri@usc.es.
50. Mason, M. G.; Botella, J. R. *J. Plant Physiol.* **1997**, *24*, 239.
51. Cabrera-Pérez, M. A.; Bermejo-Sanz, M. *Bioorg. Med. Chem.* **2004**, *12*, 5833.
52. Gia, O.; Magno, S. M.; González-Díaz, H.; Quezada, E.; Santana, L.; Uriarte, E.; DallaVia, L. *Bioorg. Med. Chem.* **2005**, *13*, 809.
53. Ewens, J. W.; Grant, R. G. *Statistical Methods in Bioinformatics: An Introduction*; Springer: New York, 2003.
54. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2004**, *12*, 745.
55. González-Díaz, H.; Ramos de Armas, R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
56. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Int. J. Mol. Sci.* **2004**, *5*, 276.
57. González-Díaz, H.; Agüero-Chapin, G.; Cabrera-Pérez, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castañedo, N. *Bioorg. Med. Chem.* **2005**, *15*, 551.
58. González-Díaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 1119.
59. Norberg, J.; Nilsson, L. *Biophys. J.* **2000**, *79*, 1537.
60. González-Díaz, H.; Marrero-Ponce, Y.; Hernández, I.; Bastida, I.; Tenorio, I.; Nasco, O.; Uriarte, E.; Castañedo, N.; Cabrera-Pérez, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
61. Dellaporta, S. L.; Word, J.; Hicks, J. B. *Plant Mol. Biol. Rep.* **1983**, *1*, 19.
62. López-Gómez, R.; Gómez-Lim, M. A. *Hortic. Sci.* **1992**, *27*, 440.