Molecular Simulation
Vol. 35, No. 3, March 2009, 220–227

Characteristic and counting polynomials: modelling nonane isomers properties

Lorentz Jäntschi[a1], Sorana D. Bolboacă[ab]* and Cristina M. Furdui[c2]

[a]*Department of Chemistry, Technical University of Cluj-Napoca, Cluj-Napoca, Romania;* [b]*Department of Medical Informatics and Biostatistics, "Iuliu Hatieganu" University of Medicine and Pharmacy Cluj-Napoca, Cluj-Napoca, Romania;* [c]*Department of Molecular Medicine, Wake Forest University Health Sciences, Winston-Salem, NC 27157, USA*

(*Received 2 May 2008; final version received 10 August 2008*)

The major goal of this study was to investigate the broad application of graph polynomials to the analysis of Henry's law constants (solubility) of nonane isomers. In this context, Henry's law constants of nonane isomers were modelled using characteristic and counting polynomials. The characteristic and counting polynomials on the distance matrix (CDi), on the maximal fragments matrix (CMx), on the complement of maximal fragments matrix (CcM) and on the Szeged matrix (CSz) were calculated for each compound. One of the nonane isomers, 4-methyloctane, was identified as an outlier and was withdrawn from further analysis. This report describes the performance and characteristics of most significant models. The results showed that Henry's law constants of nonane isomers could be modelled by using characteristic polynomial and counting polynomial on the distance matrix.

Keywords: characteristic polynomial; counting polynomials; nonane isomers; Henry's law constant (solubility)

1. Introduction

Computational methods are being used today for the characterisation of chemical compounds and to get a better understanding of the relationships between their structure and physical, chemical and/or biological properties.

The characteristic polynomial is defined in algebra as a polynomial associated to any square matrix [1]. The characteristic polynomial encodes several properties of a matrix, the most important being the matrix eigenvalues, its determinant and its trace [2]. A characteristic polynomial can be defined as:

$$\varphi(G, X) = \det[X\mathrm{I} - A(G)], \tag{1}$$

where $A(G)$ is the adjacency matrix of a pertinently constructed skeleton graph and I is the identity matrix [3].

Many studies were reported on the application of characteristic polynomials in different research fields such as mathematics [4,5], computer science [6–8], engineering [9], chemistry [10–12], physics [13,14] and management [15]. The characteristic polynomial is the most popular and the most extensively studied graph polynomial in chemical graph theory [3]. The characteristic polynomials proved their performances in correlations as molecular descriptors in the characterisation of the properties of chemical compounds [16].

Counting polynomials are also used in chemical graph theory. The general formula of a counting polynomial is:

$$\sum_{k\geq 0} a_k X^k, \quad \text{where } a_k = |\{M_{i,j} | M_{i,j} = k\}|, \tag{2}$$

a_k being the polynomial-count and $i, j = 1, \ldots, n$.

Some methods that use the distance matrix, the Szeged matrix or the Cluj matrix were reported in literature as methods for counting polynomials [3].

Solvation is extremely important, because the large majority of (bio)chemical processes takes place in the liquid phase. Solvation free energies (free energies for the transfer of solute from the gas phase to solution) can be calculated by quantum chemical methods in conjunction with implicit solvent models like solvent reaction field [17,18] and Langevin dipoles [19,20] or by molecular dynamics simulations in conjunction with explicit solvent and free energy perturbation [21]. However, since such calculations are extremely time consuming, there exists an urgent need for development of simpler approaches to accurately predict solvation free energies.

The aim of this study was to analyse the Henry's law constant (solubility) of nonane isomers by using characteristic and counting polynomials and to prove that characteristic and counting polynomials can be used to characterise the relationship between structure and chemical properties for this class of compounds.

*Corresponding author. Email: sbolboaca@umfcluj.ro

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020802398892
http://www.informaworld.com

2. Materials and methods

Alkanes are acyclic saturated hydrocarbon structures that normally have a linear configuration. The general chemical formula is C_nH_{2n+2}. It is well known that the number of isomers increases with the number of carbon atoms, for the alkanes with 1–10 carbons, the number of isomers being equal with 1, 1, 1, 2, 3, 5, 9, 18, 35, and 75, respectively.

This study focuses on nonane isomers with the general chemical structure C_9H_{20}. The systematic names of the compounds studied are: 4-methyloctane (a_01), 3-ethyl-2,3-dimethylpentane (a_02), 3,3-diethylpentane (a_03), 2,2,3,3-tetramethyl-pentane (a_04), 2,3,3,4-tetramethyl-pentane (a_05), nonane (a_06), 2,3,3-trimethylhexane (a_07), 3,3,4-trimethylhexane (a_08), 3-ethyl-3-methyl-hexane (a_09), 2,2,3,4-tetra-methylpentane (a_10), 3,4-dimethylheptane (a_11), 2,3,4-trimethylhexane (a_12), 3-ethyl-4-methylhexane (a_13), 3-ethyl-2,2-dimethylpentane (a_14), 3-ethyl-2,4-dimethylpentane (a_15), 2,3-dimethylheptane (a_16), 3,3-dimethylheptane (a_17), 4,4-dimethylheptane (a_18), 3-ethylheptane (a_19), 4-ethyl-heptane (a_20), 2,2,3-trimethylhexane (a_21), 2,2,5-trimethylhexane (a_22), 2,4,4-trimethylhexane (a_23), 3-ethyl-2-methylhexane (a_24), 2,2,4,4-tetramethylpentane (a_25), 3-methyloctane (a_26), 2,5-dimethylheptane (a_27), 3,5-dimethyl-heptane (a_28), 2,3,5-trimethylhexane (a_29), 2-methyloctane (a_30), 2,2-dimethylheptane (a_31), 2,4-dimethylheptane (a_32), 2,6-dimethylheptane (a_33), 2,2,4-trimethyl-hexane (a_34) and 4-ethyl-2-methyl-hexane (a_35), respectively.

The Henry's law constant (solubility of a gas in water) of alkanes expressed as trace gases of potential importance in environmental chemistry was the property of interest. The measured values were taken from previously reported research [22] (k_H, Table 1) and were given as M/atm unit measurements (M/atm = [mol_{aq}/dm^3_{aq}]/atm).

The Henry's law constant was modelled by using characteristic and counting polynomials (Equations (1) and (2), respectively). Four matrices were used for counting polynomials: the distance matrix (CDi), the maximal fragments matrix (CMx), the complement of the maximal fragments matrix (CcM) and the Szeged matrix (CSz) [23,24].

A monovariate model based on characteristic polynomials was constructed in order to identify the outliers.

The correlation coefficient between measured and estimated values by the model greater than 0.2 (even if it is well known that a value less than 0.25 indicate the absence of a linear relationship [25]) was the criterion imposed in identification of the characteristic and counting polynomials models. The multivariate models were obtained by using homemade software that implemented a systematic search using rational numbers (p/q) as roots based on the imposed criterion:$-100 \leq p, q \leq 100$. For the models with good estimated ability ($r > 0.75$ [26]; r, correlation coefficient) a systematic search was applied for $0 < p, q \leq 50$ considering the whole sample of 35 and sampled obtained by excluding the outliers (if any exists).

The methodology applied to assess the validity and reliability of the identified polynomials models was as follows:

- Step 1: leave-one-out cross-validation analysis. The techniques employed a number of training sets equal to the number of investigated molecules minus one, and from each of these samples one compound was excluded. A model was obtained and it was used to predict the property of excluded compound for each training set.
- Step 2: leave-n%-out cross-validation as internal validation analysis (n, valid sample size). A number of 1/3 from the total number of compound in the sample was randomly chosen to be included into the test set. The remained compounds were used to build the model; the model was applied on test set. The model obtained in test set was considered valid and stable when the correlation coefficient on test set was not statistically different by the correlation coefficient on training set.
- Step 3: leave-n%-out cross-validation as external validation analysis. The sample of valid compounds (excluding the outliers if any exists) was randomly split into training and test set. One-third of compounds were included into test set. The training set compound were used in order to identify the characteristics and counting polynomial on different matrixes according to the abilities obtained when all compounds were investigated. The criterion used in roots search was $-100 \leq p, q \leq 100$. The obtained model with higher abilities in estimation was used in order to predict the property on test set. The correlation coefficient and associated 95% confidence interval (95%CI), the Fisher parameter and associated significance were used in order to validate the model, on both training and test set.
- Step 4: correlation coefficient comparison analysis between and within models. The correlation coefficients obtained by different models were then analysed and compared using the Steiger's Z-test [27] at a significance level of 5%.

3. Results and discussion

The characteristic polynomial (ChP) and CDi, CMx, CcM, CSz counting polynomials were calculated for each nonane.

To determine the irreducible or primer factors, the characteristic and counting polynomials obtained as

Table 1. Characteristic and counting polynomials for nonane isomers: the values of the $Q(X)$ terms.

No.	$k_H(\times 10^5)$ [M/atm]	$Q(X)_{ChP}$	$Q(X)_{CDi}$	$Q(X)_{CMx}$	$Q(X)_{CcM}$	$Q(X)_{CSz}$
a_01	1.0	$(2X-1)(2X+1)(5X^2-3)$	$X^5+2X^4+4X^3+6X^2+7X+8$	$8X^7+14X^6+12X^5+5X^4+4X^3+6X^2+4X+1$	$X^7+4X^6+6X^5+4X^4+5X^3+12X^2+14X+8$	$X^7+6X^6+13X^5+9X^4+9X^3+16X^2+9X+3$
a_02	1.5	$17X^4-12X^2+2$	$5X^2+12X+11$	$24X^7+14X^6+6X^5+3X^2+4X+3$	$3X^7+4X^6+3X^5+6X^2+14X+24$	$3X^7+10X^6+9X^5+14X^2+20X+10$
a_03	1.5	$18X^4-16X^2+5$	$2(3X^2+6X+5)$	$2(8X^7+14X^6+4X+1)$	$2(X^7+4X^6+14X+8)$	$2(X^7+10X^6+20X+2)$
a_04	1.6	$3X^2(5X^2-2)$	$3X^2+12X+13)$	$32X^7+7X^6+5X^4+4X^3+2X+4$	$4X^7+2X^6+4X^4+5X^3+7X+32$	$4X^7+5X^6+13X^4+16X^3+10X+18$
a_05	1.6	$8X^2(2X^2-1)$	$4(X^2+3X+3)$	$2(16X^7+6X^5+3X^2+2)$	$2(2X^7+3X^5+6X^2+16)$	$2(2X^7+9X^5+14X^2+8)$
a_06	1.7	$21X^4-20X^2+5$	$X^6+2X^5+3X^4+4X^3+5X^2+6X+7$	$2X(7X^5+6X^4+5X^3+4X^2+3X+2)$	$2X(2X^5+3X^4+4X^3+5X^2+6X+7)$	$2X(X^2+X+1)(3X^3+2X^2+2X+4)$
a_07	1.7	$X^2(17X^2-10)$	$2X^3+5X^2+10X+11$	$24X^7+7X^6+12X^5+6X^2+2X+3$	$3X^7+2X^6+6X^5+12X^2+7X+24$	$3X^7+3X^6+19X^5+23X^2+6X+12$
a_08	1.7	$17X^4-11X^2+2$	$X^3+5X^2+11X+11$	$24X^7+14X^6+5X^4+4X^3+4X+3$	$3X^7+4X^6+4X^4+5X^3+14X+24$	$(X+1)(3X^6+6X^5-6X^4+19X^3-6X^2+6X+11)$
a_09	1.7	$18X^4-14X^2+3$	$2(X^3+3X^2+5X+5)$	$16X^7+21X^6+6X^5+3X^2+6X+2$	$2X^7+6X^6+3X^5+6X^2+21X+16$	$2X^7+13X^6+10X^5+11X^2+24X+6$
a_10	1.7	$2X^2(8X^2-3)$	$2(3X^2+5X+6)$	$32X^7+6X^5+5X^4+4X^3+3X^2+4$	$4X^7+3X^5+4X^4+5X^3+6X^2+32$	$4X^7+7X^5+10X^4+15X^3+13X^2+17$
a_11	1.8	$19X^4-15X^2+3$	$(X^2+3)(X^2+3X+3)$	$16X^7+14X^6+6X^5+5X^4+4X^3+3X^2+4X+2$	$2X^7+4X^6+3X^5+4X^4+5X^3+6X^2+14X+16$	$(X+1)(2X^6+5X^5+3X^4+7X^3+4X^2+6X+6)$
a_12	1.8	$2(3X^2-1)^2$	$2(X^3+3X^2+5X+5)$	$24X^7+7X^6+6X^5+5X^4+4X^3+3X^2+2X+3$	$3X^7+2X^6+3X^5+4X^4+5X^3+6X^2+7X+24$	$3X^7+4X^6+7X^5+11X^4+13X^3+11X^2+7X+10$
a_13	1.8	$19X^4-16X^2+4$	$2X^3+7X^2+10X+9$	$16X^7+21X^6+5X^4+4X^3+6X+2$	$2X^7+6X^6+4X^4+5X^3+21X+16$	$2X^7+12X^6+11X^4+13X^3+23X+5$
a_14	1.8	$X^2(17X^2-10)$	$7X^2+10+11$	$24X^7+14X^6+5X^4+4X^3+4X+3$	$3X^7+4X^6+4X^4+5X^3+14X+24$	$3X^7+8X^6+10X^4+15X^3+18X+12$
a_15	1.8	$6X^2(3X^2-2)$	$2(4X^2+5X+5)$	$24X^7+7X^6+12X^5+6X^2+2X+3$	$3X^7+2X^6+6X^5+12X^2+7X+24$	$3X^7+4X^6+14X^5+26X^2+9X+10$
a_16	1.9	$19X^4-14X^2+2$	$2X^4+4X^3+5X^2+8X+9$	$16X^7+7X^6+12X^5+5X^4+4X^3+6X^2+2X+2$	$2X^7+2X^6+6X^5+4X^4+5X^3+12X^2+7X+16$	$2X^7+3X^6+13X^5+10X^4+9X^3+18X^2+4X+7$
a_17	1.9	$2(3X^2-1)^2$	$X^4+4X^3+5X^2+8X+10$	$16X^7+14X^6+6X^5+5X^4+4X^3+3X^2+4X+2$	$2X^7+4X^6+3X^5+4X^4+5X^3+6X^2+14X+16$	$2X^7+8X^6+7X^5+10X^4+9X^3+10X^2+12X+8$
a_18	1.9	$6X^2(3X^2-2)$	$X^4+2X^3+7X^2+8X+10$	$2(8X^7+7X^6+6X^5+3X^2+2X+1)$	$2(X^7+2X^6+3X^5+6X^2+7X+8)$	$2(X^7+3X^6+9X^5+10X^2+6X+4)$
a_19	1.9	$20X^4-18X^2+5$	$2(X^4+2X^3+3X^2+4X+4)$	$8X^7+21X^6+6X^5+5X^4+4X^3+3X^2+6X+1$	$X^7+6X^6+3X^5+4X^4+5X^3+6X^2+21X+8$	$X^7+11X^6+6X^5+10X^4+9X^3+9X^2+18X+2$
a_20	1.9	$2(2X^2-1)(5X^2-2)$	$X^4+4X^3+7X^2+8X+8$	$8X^7+21X^6+12X^5+6X^2+6X+1$	$X^7+6X^6+6X^5+12X^2+21X+8$	$X^7+10X^6+16X^5+20X^2+17X+2$
a_21	1.9	$X^2(17X^2-9)$	$3X^3+5X^2+9X+11$	$24X^7+7X^6+6X^5+5X^4+4X^3+3X^2+2X+3$	$3X^7+2X^6+3X^5+4X^4+5X^3+6X^2+7X+24$	$3X^7+3X^6+9X^5+10X^4+12X^3+11X^2+5X+13$
a_22	1.9	$X^2(17X^2-6)$	$6X^3+5X^2+6X+11$	$24X^7+6X^5+10X^4+8X^3+3X^2+3$	$3X^7+3X^5+8X^4+10X^3+6X^2+24$	$3X^7+5X^5+18X^4+20X^3+6X^2+14$
a_23	1.9	$X^2(17X^2-8)$	$2X^3+7X^2+8X+11$	$24X^7+7X^6+6X^5+5X^4+4X^3+3X^2+2X+3$	$3X^7+2X^6+3X^5+4X^4+5X^3+6X^2+7X+24$	$3X^7+5X^6+5X^5+13X^4+10X^3+10X^2+8X+12$
a_24	1.9	$19X^4-15X^2+2$	$3X^3+7X^2+9X+9$	$2(8X^7+7X^6+6X^5+3X^2+2X+1)$	$2\cdot(X^7+2\cdot X^6+3\cdot X^5+6\cdot X^2+7\cdot X+8)$	$2X^7+7X^6+16X^5+22X^2+13X+6$
a_25	1.9	$15X^4$	$9X^2+6X+13$	$2(16X^7+5X^4+4X^3+2)$	$2(2X^7+4X^4+5X^3+16)$	$2(2X^7+7X^4+14X^3+10)$
a_26	2.0	$20X^4-17X^2+4$	$X^5+3X^4+4X^3+5X^2+7X+8$	$8X^7+14X^6+6X^5+10X^4+8X^3+3X^2+4X+1$	$X^7+4X^6+3X^5+8X^4+10X^3+6X^2+14X+8$	$(X+1)(X^6+6X^5-X^4+17X^3+7X+3)$
a_27	2.0	$19X^4-13X^2+2$	$2X^4+5X^3+5X^2+7X+9$	$16X^7+7X^6+6X^5+10X^4+8X^3+3X^2+2X+2$	$2X^7+2X^6+3X^5+8X^4+10X^3+6X^2+7X+16$	$(X+1)(2X^6+2X^5+3X^4+13X^3+7X^2-X+7)$
a_28	2.0	$19X^4-14X^2+3$	$X^4+4X^3+6X^2+8X+9$	$2(8X^7+7X^6+5X^4+4X^3+2X+1)$	$2(X^7+2X^6+4X^4+5X^3+7X+8)$	$2(X^7+4X^6+8X^4+11X^3+6X+3)$
a_29	2.0	$2X^2(9X^2-5)$	$2(2X^3+3X^2+4X+5)$	$24X^7+12X^5+5X^4+4X^3+6X^2+3$	$3X^7+6X^5+4X^4+5X^3+12X^2+24$	$3X^7+12X^5+12X^4+11X^3+17X^2+11$
a_30	2.1	$2(10X^4-8X^2+1)$	$2X^5+3X^4+4X^3+5X^2+6X+8$	$8X^7+7X^6+12X^5+10X^4+8X^3+6X^2+2X+1$	$X^7+2X^6+6X^5+8X^4+10X^3+12X^2+7X+8$	$X^7+3X^6+10X^5+15X^4+17X^3+12X^2+4X+4$
a_31	2.1	$2X^2(9X^2-5)$	$3X^4+4X^3+5X^2+6X+10$	$16X^7+7X^6+6X^5+10X^4+8X^3+3X^2+2X+2$	$2X^7+2X^6+3X^5+8X^4+10X^3+6X^2+7X+16$	$2X^7+3X^6+5X^5+17X^4+17X^3+8X^2+4X+10$
a_32	2.1	$X^2(19X^2-13)$	$2X^4+3X^3+7X^2+7X+9$	$16X^7+7X^6+12X^5+5X^4+4X^3+6X^2+2X+2$	$2X^7+2X^6+6X^5+4X^4+5X^3+12X^2+7X+16$	$2X^7+3X^6+12X^5+10X^4+10X^3+17X^2+5X+7$
a_33	2.1	$X^2(19X^2-12)$	$4X^4+4X^3+5X^2+6X+9$	$2(X^2-X+1)(8X^5+8X^4+6X^3+3X^2+X+1)$	$2(X^2-X+1)(X^5+X^4+3X^3+6X^2+8X+8)$	$2(X^7+5X^5+7X^4+10X^3+6X^2+4)$
a_34	2.1	$X^1(17X^2-7)$	$3X^3+7X^2+7X+11$	$24X^7+7X^6+10X^4+8X^3+2X+3$	$3X^7+2X^6+8X^4+10X^3+7X+24$	$(X+1)(3X^6+X^5-X^4+17X^3+7X^2-7X+13)$
a_35	2.1	$19X^4-14X^2+2$	$4X^3+7X^2+8X+9$	$16X^7+14X^6+6X^5+5X^4+4X^3+3X^2+4X+2$	$2X^7+4X^6+3X^5+4X^4+5X^3+6X^2+14X+16$	$2X^7+8X^6+5X^5+12X^4+11X^3+8X^2+14X+6$

a_01, 4-methyloctane; a_02, 3-ethyl-2,3-dimethylpentane; a_03, 3,3-diethylpentane; a_04, 2,2,3,3-tetramethyl-pentane; a_05, 2,3,3,4-tetramethylpentane; a_06, nonane; a_07, 2,3,3-trimethylhexane; a_08, 3,3,4-trimethylhexane; a_09, 3-ethyl-3-methylhexane; a_10, 2,2,3,4-tetra-methylpentane; a_11, 3,4-dimethylheptane; a_12, 2,3,4-trimethylhexane; a_13, 3-ethyl-4-methylhexane; a_14, 3-ethyl-2,2-dimethylpentane; a_15, 3-ethyl-2,4-dimethylpentane; a_16, 2,3-dimethylheptane; a_17, 3,3-dimethylheptane; a_18, 4,4-dimethylheptane; a_19, 3-ethylheptane; a_20, 4-ethyl-heptane; a_21, 2,2,3-trimethylhexane; a_22, 2,2,5-trimethylhexane; a_23, 2,4,4-trimethylhexane; a_24, 3-ethyl-2-methylhexane; a_25, 2,2,4,4-tetramethylpentane; a_26, 3-methyloctane; a_27, 2,5-dimethylheptane; a_28, 3,5-dimethyl-heptane; a_29, 2,3,5-trimethylhexane; a_30, 2-methyloctane; a_31, 2,2-dimethylheptane; a_32, 2,4-dimethylheptane; a_33, 2,6-dimethylheptane; a_34, 2,2,4-trimethyl-hexane; a_35 and 4-ethyl-2-methyl-hexane.

described above were factorised. The generic formulas are described below:

- ChP:

$$P(X)_{\mathrm{ChP}} = X^7(X^2 - 8) + X{\cdot}Q(X)_{\mathrm{ChP}}, \tag{3}$$

- CDi:

$$P(X)_{\mathrm{CDi}} = 2X^2Q(X)_{\mathrm{CDi}} + 16X + 9, \tag{4}$$

- CMx:

$$P(X)_{\mathrm{CMx}} = 16X^8 + XQ(X)_{\mathrm{CMx}} + 2X + 9, \tag{5}$$

- CcM:

$$P(X)_{\mathrm{CcM}} = 2X^8 + XQ(X)_{\mathrm{CcM}} + 16X + 9, \tag{6}$$

- CSz:

$$P(X)_{\mathrm{CSz}} = 2X^8 + XQ(X)_{\mathrm{CSz}} + 4X + 9, \tag{7}$$

where $P(X)_{\mathrm{ChP}}$, $P(X)_{\mathrm{CDi}}$, $P(X)_{\mathrm{CMx}}$, $P(X)_{\mathrm{CcM}}$, $P(X)_{\mathrm{CSz}}$ are the characteristic polynomial and counting polynomials on the: CDi, CMx, CcM and on the Szeged matrix, respectively. The $Q(X)$ values for each type of polynomial are presented in Table 1.

By analysing the polynomials described above, it can be observed that the characteristic polynomial (Equation (3)) can be easily factorised while the counting polynomials (Equations (4)–(7)) are not. The characteristic polynomial includes other invariants called characteristic solutions and this could explain the observation above.

Regarding the formulas obtained for counting polynomials (Equations (4)–(7)) the following similarities can be observed:

(1) All formulas contain the "a_1X+9", where a_1 varies from 2 to 16, but is always an even number. The generic formula for CMx, CcM and CSz counting polynomials is: $P(X) = a_0X^8 + XQ(X) + a_1X + 9$, where a_0 and a_1 are even integers with values from two to sixteen;

(2) The term $Q(X)$ could be factorised in limited cases (see Table 1).

The monovariate model obtained using the characteristic polynomial was:

$$\hat{Y}_{\mathrm{ChP-mono}} = 19.54 + 0.17{\cdot}P(2923/1725), \tag{8}$$

where $\hat{Y}_{\text{ChP-mono}}$ is the characteristic polynomial. A square correlation coefficient of 0.2968 was obtained for model (8) when all compounds were included and 0.6301 when the compound 4-methyloctane was withdrawn. Therefore, compound 4-methyloctane was considered an outlier and was excluded from further analysis.

There could not be identified any valid model by using neither CMx nor the CcM.

A number of valid models were obtained using the characteristic polynomial (ChP) and the counting polynomials on the distance matrix (CDi) and on the Szeged matrix on the sample of 34 compounds (more than one model with the same value of determination coefficient). As justified above, the 4-methyloctane was considered outlier and was not included in analysis. The measured value of the Henry's law constant of the excluded compound had a lower value comparing with the rest of compounds (see Table 1); this means that an error could have occurred during the experimental process.

- Characteristic polynomial:

$$\begin{aligned}\hat{Y}_{\mathrm{ChP}} = {} & -58.11 - 329.00{\cdot}P(1/100) \\ & + 8.39{\cdot}P(35/97) + 7.81 \\ & \times 10^{-3}{\cdot}P(72/23),\end{aligned} \tag{9}$$

  where $P(X_{\mathrm{i}})$ are the characteristic polynomials.

- Counting polynomial on the distance matrix:

$$\begin{aligned}\hat{Y}_{\mathrm{CDi}} = {} & 142.20 + 5.70{\cdot}P(-23/71) \\ & - 10.00{\cdot}P(5/18) - 2.11 \\ & \times 10^{-8}{\cdot}P(99/10),\end{aligned} \tag{10}$$

  where $P(X_{\mathrm{i}})$ are the CDi.

- Counting polynomial on the Szeged matrix:

$$\begin{aligned}\hat{Y}_{\mathrm{CSz}} = {} & -34.39 + 0.04{\cdot}P(-29/39) \\ & + 1.19{\cdot}P(11/9) - 0.64{\cdot}P(59/45),\end{aligned} \tag{11}$$

  where $P(X_{\mathrm{i}})$ are counting polynomials on the Szeged matrix.

The analysis of the models described above was performed by calculating the correlation coefficient (r) and the associated 95% confidence intervals (95%CI$_{\mathrm{r}}$), the standard error of the estimated (SErr) and the Fisher parameter of the model (F) and its significance for the sample size of 34 compounds. The above parameters and the confidence intervals for intercept and polynomial coefficients used by Equations (9)–(11) are presented in Table 2.

All models described by Equations (9)–(11) were statistically significant (the probability associated to the wrong model less than 0.001, see Table 2). The analysis of the correlation coefficients and associated 95%CI leads to the conclusion that the best model is the one described

Table 2. Statistical characteristics of the models from Equations (9) to (11) ($n = 34$).

		Parameter			Model coefficients			
Model	Pol	r [95%CI$_r$]	SErr	F (p)	Intercept lower Intercept upper	Desc1 lower Desc1 upper	Desc2 lower Desc2 upper	Desc3 lower/Desc3 upper
Equation (9)	ChP	0.8690 [0.7517–0.9329]	0.89	31 2.68×10^{-9}	–75.6–40.55	−462.44–195.56	4.10 12.67	5.99×10^{-3} 9.62×10^{-3}
Equation (10)	CDi	0.9239 [0.8518–0.9616]	0.69	58 1.28×10^{-12}	122.83 161.57	4.31 7.08	−11.56 −8.44	-2.82×10^{-8}–1.39×10^{-8}
Equation (11)	CSz	0.7188 [0.5028–0.8502]	1.26	11 6.08×10^{-5}	–55.94–12.84	−0.04 0.13	0.73 1.65	−0.92−0.36

r, correlation coefficient; 95%CI$_r$, 95% confidence intervals for correlation coefficient; SErr, standard error of estimated; *F*, Fisher parameter; Intercept lower, the lower border of the 95% confidence interval for intercept; Intercept upper, the upper border of the 95% confidence interval for intercept; Desc1 = ChP(1/100) – Equation (9); CDi(−23/71) – Equation (10); CDi(−29/39) – Equation (11); Desc2 = ChP(35/97) – Equation (9); CDi(5/18) – Equation (10); CDi(11/9) – Equation (11); Desc3 = ChP(89/18) – Equation (9); CDi(99/10) – Equation (10); CDi(59/45) – Equation (11)

by the counting polynomial on the distance matrix presented in Equation (10). Eighty-five percent of the Henry's law constant variation of the nonane isomers included in this study can be explained by its linear relationship with the variation of counting polynomial on the distance matrix used in the model.

Two counting polynomials models revealed to have estimated abilities, counting polynomial on the distance matrix and on the Szeged matrix. The difference between the correlation coefficient obtained by Equations (10) and (11) was of 0.2051. Almost 53% of the Henry's law constant variation of studied nonane isomers can be explained by its linear relationship with the variation of counting polynomial on the Szeged matrix.

The analysis of the correlation coefficients and their associated 95%CI showed that there are no significant differences between models from Equations (9) and (10) or between models in Equations (9)–(11), respectively, due to the existence of the overlap of those intervals.

The results obtained in leave-one-out internal validation analysis (see Table 3) showed a difference between correlation coefficient of the model and correlation coefficient obtained in leave-one-out analysis of 0.05 for the models from Equations (9) and (10), and of 0.12 for the model from Equation (11). These results sustain the stability of the models from Equations (9)–(11) [25].

Steiger's *Z*-test was then used to test the hypothesis that there were no significant differences between correlation coefficients obtained by models from Equations (9)–(11). The matrix of *p*-values associated to the *Z* parameters is presented in Table 4. The results revealed that the models from Equations (9) and (10) had the same ability in estimates of the relationship between nonane isomers structure and property of interest.

The ability of the models from Equations (9) and (10) was investigated by applying the following systematic search $0 < p,\ q \leq 50$. Three sample sizes were considered: 35 (all compounds), 34 (excluding the 4-methyloctane compound that proved to be an outlier) and 33 (excluding the 4-methyloctane and nonane, nonane seems to be an outlier if the distribution of the measured property is analysed), respectively.

The models from Equations (12) to (14) were obtained when the characteristic polynomial was investigated:

$$\hat{Y}_{\mathrm{ChP}} = -7828.32 - 435.57 \cdot \mathrm{P}(1/50) + 33.31 \cdot \mathrm{P}(12/47) + 7.75 \times 10^{-4} \cdot \mathrm{P}(34/7), \quad (12)$$

where $\mathrm{P}(X_i)$ are the characteristic polynomials. Statistical characteristics of the models are as follows: $r = 0.5884$, 95%CI$_r$ [0.3173–0.7705], SErr = 1.89, $F = 5$, $p = 3.89 \times 10^{-3}$; $r_{\mathrm{loo}} = 0.4692$ (correlation coefficient obtained in leave-one-out cross-validation analysis), $F_{\mathrm{loo}} = 3$ (Fisher parameter obtained in leave-one-out

Table 3. Leave-one-out cross validation results.

Model	r_{pred} [95%CI]	$SErr_{pred}$	F_{pred} (p_{pred})
Equation (9)	0.8206 [0.6644–0.9080]	1.04	21 (2.43×10^{-7})
Equation (10)	0.8714 [0.7534–0.9349]	0.89	31 (2.7×10^{-9})
Equation (11)	0.6008 [0.3243–0.7826]	1.47	5 (5.15×10^{-3})

r, correlation coefficient; 95%CI, 95% confidence interval of *r*; SErr, standard error of predicted; F_{pred}, Fisher parameter in leave-one-out analysis; p_{pred}, probability associated to F_{pred}

cross-validation analysis), $p_{loo} = 7.51 \times 10^{-2}$ (significance of the model obtained in leave-one out analysis).

$$\hat{Y}_{ChP} = -1683.73 - 441.75 \cdot P(1/50) + 33.65 \cdot P(12/47) + 4.46 \times 10^{-4} \cdot P(50/9), \quad (13)$$

where $P(X_i)$ are the characteristic polynomials. Statistical characteristics of the models are as follows: $r = 0.8690$, $95\%CI_r$ [0.7517–0.9329], SErr = 0.89, $F = 31$, $p = 2.68 \times 10^{-9}$; $r_{loo} = 0.8206$ (correlation coefficient obtained in leave-one-out cross-validation analysis), $F_{loo} = 20$ (Fisher parameter obtained in leave-one-out cross-validation analysis), $p_{loo} = 2.43 \times 10^{-7}$ (significance of the model obtained in leave-one out analysis). Note that there were identified a number of 1252 models that had a determination coefficient of 0.755.

$$\hat{Y}_{ChP} = 20.21 - 117.95 \cdot P(1/50) + 8.40 \cdot P(1/4) + 5.28 \times 10^{-2} \cdot P(50/23), \quad (14)$$

where $P(X_i)$ are the characteristic polynomials. Statistical characteristics of the models are as follows: $r = 0.9194$, $95\%CI_r$ [0.8417–0.9597], SErr = 0.71, $F = 53$, $p = 7.16 \times 10^{-12}$; $r_{loo} = 0.8958$ (correlation coefficient obtained in leave-one-out cross-validation analysis), $F_{loo} = 39$ (Fisher parameter obtained in leave-one-out cross-validation analysis), $p_{loo} = 2.74 \times 10^{-10}$ (significance of the model obtained in leave-one out analysis). Note that a number of 1352 models had a determination coefficient of 0.845.

The analysis of Equations (12)–(14) revealed the followings:

- Even if the model from Equation (12) is statistically significant and the correlation coefficient in test set is included into the 95%CI of the correlation coefficient obtained in training sent, the model in test set is not statistically significant;
- A significant increase of correlation coefficient is observed when the 4-methyloctane compound is excluded, proving that it is an outlier;
- A determination of 85% is obtained when both 4-methyloctane and nonane are excluded from the sample when the best model is searched. This suggests that the nonane compound could be also an outlier.

Table 4. Correlated correlation analysis: Steiger's *Z*-test applied on Equations (9)–(11).

Models	Steiger *Z* parameter	*p*-value
Equations (9)–(10)	–1.6794	9.53×10^{-1}
Equations (9)–(11)	2.84439	2.22×10^{-3}
Equations (10)–(11)	3.53456	2.04×10^{-4}

The following models were obtained by investigation of counting polynomial on the distance matrix:

$$\hat{Y}_{CDi} = -19.74 - 0.15 \cdot P(-35/36) + 0.29 \cdot P(19/13) + 6.53 \times 10^{-2} \cdot P(43/24), \quad (15)$$

where $P(X_i)$ are the counting polynomial on the distance matrix. Statistical characteristics of the models are as follows: $r = 0.5912$, $95\%CI_r$ [0.3212–0.7722], SErr = 1.89, $F = 6$, $p = 3.61 \times 10^{-3}$; $r_{loo} = 0.4512$ (correlation coefficient obtained in leave-one-out cross-validation analysis), $F_{loo} = 2$ (Fisher parameter obtained in leave-one-out cross-validation analysis), $p_{loo} = 1.74 \times 10^{-1}$ (significance of the model obtained in leave-one out analysis).

$$\hat{Y}_{CDi} = 152.42 - 6.28 \cdot P(-11/35) - 11.16 \cdot P(4/15) + 2.28 \times 10^{-8} \cdot P(49/5), \quad (16)$$

where $P(X_i)$ are the counting polynomial on the distance matrix. Statistical characteristics of the models are as follows: $r = 0.9239$, $95\%CI_r$ [0.8518–0.9616], SErr = 0.69, $F = 58$, $p = 1.27 \times 10^{-12}$; $r_{loo} = 0.8844$ (correlation coefficient obtained in leave-one-out cross-validation analysis), $F_{loo} = 35$ (Fisher parameter obtained in leave-one-out cross-validation analysis), $p_{loo} = 6.05 \times 10^{-10}$ (significance of the model obtained in leave-one out analysis). Note that there were identified a number of 563 models that had a determination coefficient of 0.854.

$$\hat{Y}_{CDi} = 39.33 + 13.19 \cdot P(-11/47) - 9.80 \cdot P(15/32) + 3.31 \cdot P(24/35), \quad (17)$$

where $P(X_i)$ are the counting polynomial on the distance matrix. Statistical characteristics of the models are as follows: $r = 0.9234$, $95\%\mathrm{CI}_r$ [0.8493–0.9617], SErr = 0.70, $F = 56, p = 3.54 \times 10^{-12}$; $r_{\mathrm{loo}} = 0.8873$ (correlation coefficient obtained in leave-one-out cross-validation analysis), $F_{\mathrm{loo}} = 34$ (Fisher parameter obtained in leave-one-out cross-validation analysis), $p_{\mathrm{loo}} = 9.34 \times 10^{-10}$ (significance of the model obtained in leave-one out analysis). Note that a number of 6095 models had a determination coefficient of 0.853.

The analysis of the models from Equations (15)–(17) revealed the followings:

(1) The model obtained by Equation (15) for test set is not statistically significant.
(2) The models from Equations (16) and (17) are almost identical in terms of correlation coefficients in both training and test sets, standard error of estimated, Fisher parameters and associated significances.
(3) A determination of 85% is obtained both when 4-methyloctane, respectively, 4-methyloctane and nonane are excluded from the sample. The last observation suggested that the nonane compound could be also an outlier. In these conditions, the nonane could not be considered as an outlier.

The external validation analysis was performed in order to validate the contribution of the characteristic and counting polynomial on the distance matrix in characterisation of the relationship between nonane isomers' structure and Henry's law constant. The following compounds were assigned randomly into test set: 3-ethyl-2,3-dimethylpentane (a_02), 3-methyloctane (a_26), 2,3,3-trimethylhexane (a_07), 3,3-dimethylheptane (a_17), 3,3-diethylpentane (a_03), 2-methyloctane (a_30), 2,5-dimethylheptane (a_27), nonane (a_06), 4-ethyl-heptane (a_20), 2,2,3,4-tetra-methylpentane (a_10) and 2,3,3,4-tetramethylpentane (a_05).

A characteristic polynomial model obtained in external validation analysis is presented in Equation (18):

$$\hat{Y}_{\mathrm{ChP}} = -63.06 - 218.52 \cdot P(1/100) + 4.72 \cdot P(2/5) + 6.18 \times 10^{-3} \cdot P(77/24), \tag{18}$$

where $P(X_i)$ are the characteristic polynomials. Statistical characteristics of the models are as follows: $r_{\mathrm{tr}} = 0.9092$ (correlation coefficient in training set), $95\%\mathrm{CI}_{\mathrm{rtr}}$ [0.7948–0.9611], $\mathrm{SErr}_{\mathrm{tr}} = 0.63$ (standard error of estimated), $F_{\mathrm{tr}} = 30$ (Fisher parameter in training set), $p = 1.95 \times 10^{-7}$; $r_{\mathrm{ts}} = 0.8042$ (correlation coefficient in test set), $F_{\mathrm{ts}} = 16$ (Fisher parameter in test set), $p_{\mathrm{ts}} = 2.84 \times 10^{-3}$ (significance of the F_{ts}). A number of 2045 models that have a determination coefficient of 0.853 were obtained. The external validation analysis revealed that the characteristic polynomial leads to a valid and reliable solution in characterisation of the relationship between the structure of nonane isomers and property of interest providing good models with abilities in estimation as well as in prediction. The correlation coefficient obtained in external validation on both training and test set is not statistically significant different by the one provided by Equation (9) (both correlation coefficients are included into the 95%CI of correlation coefficient of model from Equation (9)).

A counting polynomial on distance matrix model is presented in Equation (19):

$$\hat{Y}_{\mathrm{CDi}} = 95.07 + 1.76 \cdot P(-49/97) - 4.94 \cdot P(32/99) + 1.77 \times 10^{-7} \cdot P(22/5), \tag{19}$$

where $P(X_i)$ are the counting polynomial on the distance matrix. Statistical characteristics of the models are as follows: $r = 0.9058$, $95\%\mathrm{CI}_r$ [0.7876–0.9596], SErr = 0.64, $F = 29$, $p = 2.70 \times 10^{-7}$; $r_{\mathrm{ts}} = 0.7429$ (correlation coefficient in test set), $F_{\mathrm{ts}} = 11$ (Fisher parameter for test set), $p_{\mathrm{ts}} = 8.80 \times 10^{-3}$ (significance of the model in test set). Note that a number of 5001 models had a determination coefficient of 0.821. The analysis of the results obtained by counting polynomial on distance matrix revealed that the correlation coefficient obtained in test set is not contained into the 95%CI of the correlation coefficient in training set. This suggested that there is a statistically significant difference between estimated and prediction ability of the model from Equation (19). The comparison of those correlation coefficients leads to a probability of a type I error of 0.096 ($Z = 1.307$), sustaining that there is no statistically significant differences between them.

The Steiger Z-test was applied in order to identify if there is a significant difference between correlation coefficient obtained by Equation (18) and the one obtained by Equation (19). The obtained Z-score of 0.1211 ($p = 4.52 \times 10^{-1}$) leads to the conclusion that both characteristic polynomial and counting polynomial on distance matrix had good abilities in characterisation of the link between nonane isomers structure and property of interest.

No significant differences were identified between correlation coefficients obtained by the characteristic polynomial (Equations (9) and (18)) and by the counting polynomial on the distance matrix (Equations (10) and (19)). Thus, it can be concluded that there are no differences between characteristic polynomial model and counting polynomial on the distance matrix model, these two polynomials being considered useful in characterisation of the relationship between structure and property of interest on the investigated sample. There could not be identified any model with estimated ability when CMx and on the CcM were investigated. The model obtained by using the counting polynomial on the Szeged matrix proved to have significantly lower performances compared with characteristic polynomial and CDi in characterisation of the link between structure of nonane isomers and investigated property.

The aim of the research was to model the Henry's law constant by using characteristic and counting polynomials and the results showed that this is a feasible approach when characteristic polynomial or counting polynomial on distance matrix are used. The results of this study constitute a novel direction in the analysis and characterisation of chemical compounds by using mathematical models. The broad application of characteristic and counting polynomials in modelling nonane isomers properties will be investigated by modelling other physical and chemical properties of these compounds.

4. Conclusions

The Henry's law constant of the nonane isomers can be modelled using characteristic polynomial and counting polynomial on the distance matrix. These polynomials provided reliable and valid models, opening a new venue for the characterisation of chemical compounds.

Current research in our laboratory is focused on the characterisation of other properties and/or other chemical compounds to test the usefulness of the characteristic and counting polynomials in investigation of the structure–property/activity relationships.

Acknowledgements

The research was partly supported by UEFISCSU Romania through project ET108/2006.

Notes

1. Email: lori@chimie.utcluj.ro
2. Email: cfurdui@wfubmc.edu

References

[1] N. Trinajstić, *Chemical Graph Theory*, 2nd ed., CRC Press, Boca Raton, FL, 1983, Revised.

[2] N. Trinajstić, *The characteristic polynomial of a chemical graph*, J. Math. Chem. 2 (1988), pp. 197–215.

[3] M.V. Diudea, I. Gutman, and L. Jäntschi, *Molecular Topology*, 2nd ed., Nova Science, Huntington, NY, 2002, pp. 53–100.

[4] E. Strahov and Y.V. Fyodorov, *Universal results for correlations of characteristic polynomials: Riemann-Hilbert approach*, Comm. Math. Phys. 241 (2003), pp. 343–382.

[5] V.N. Kublanovskaya, *Solution of spectral problems for polynomial matrices*, J. Math. Sci. 127 (2005), pp. 2024–2032.

[6] J. Abdeljaoued and G.I. Malaschonok, *Efficient algorithms for computing the characteristic polynomial in a domain*, J. Pure Appl. Algebra 156 (2001), pp. 127–145.

[7] G.J. Lastman and N.K. Sinha, *Robust stability of discrete-time systems*, Int. J. Syst. Sci. 30 (1999), pp. 451–453.

[8] E. Kaltofen and G. Villard, *On the complexity of computing determinants*, Comput. Complexity 13 (2005), pp. 91–130.

[9] W. Tang and J. Kang, *Characteristic polynomial assignment in F–M model II of 2-D systems*, J. Syst. Eng. Electron. 15 (2004), pp. 533–536.

[10] A.T. Balaban and F. Harary, *The characteristic polynomial does not uniquely determine the topology of a molecule*, J. Chem. Documentation 11 (1971), pp. 258–259.

[11] M. Kunz, *A note on Cluj weighted adjacency matrices*, J. Serb. Chem. Soc. 63 (1998), pp. 647–652.

[12] J.R. Dias, *Properties and relationships of conjugated polyenes having a reciprocal eigenvalue spectrum – Dendralene and radialene hydrocarbons*, Croat. Chem. Acta 77 (2004), pp. 325–330.

[13] E. Breézin and S. Hikami, *Characteristic polynomials of random matrices at edge singularities*, Phys. Rev. E 62 (2000), pp. 3558–3567.

[14] E.N. Gryazina, *The D-decomposition theory*, Automat. Rem. Contr. 65 (2004), pp. 1872–1884.

[15] H. Zhang, G. Huang, and W. Zhou, *Condition of applying the fourth order of characteristic equation to the dynamic stability of wing-in-ground effect vehicles*, J. Shanghai Jiao. Univ. 34 (2000), pp. 80–82.

[16] S.D. Bolboacă and L. Jäntschi, *How good the characteristic polynomial can be for correlations?* Int. J. Mol. Sci. 8 (2007), pp. 335–345.

[17] U. Bren, M. Zupan, F.P. Guengerich, and J. Mavri, *Chemical reactivity as a tool to study carcinogenicity: Reaction between chloroethylene oxide and guanine*, J. Org. Chem. 71 (2006), pp. 4078–4084.

[18] S. Mierts, E. Scrocco, and J. Tomasi, *Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects*, Chem. Phys. 55 (1981), pp. 117–129.

[19] J. Florián and A. Warshel, *Langevin dipoles model for* ab initio *calculations of chemical processes in solution: Parametrization and application to hydration free energies of neutral and ionic solutes and conformational analysis in aqueous solution*, J. Phys. Chem. B 101 (1997), pp. 5583–5595.

[20] U. Bren, F.P. Guengerich, and J. Mavri, *Guanine alkylation by the potent carcinogen aflatoxin B1: Quantum chemical calculations*, Chem. Res. Toxicol. 20 (2007), pp. 1134–1140.

[21] U. Bren, V. Martínek, and J. Florián, *Decomposition of the solvation free energies of deoxyribonucleoside triphosphates using the free energy perturbation method*, J. Phys. Chem. B 110 (2006), pp. 12782–12788.

[22] C.L. Yaws and H.-C. Yang, *Henry's law constant for compound in water*, in *Thermodynamic and Physical Property Data*, C. L. Yaws ed., Gulf Publishing Company, Houston, TX, 1992, pp. 181–206.

[23] L. Jäntschi and S.D. Bolboacă, *Counting polynomials on regular iterative structures*, *Entropy*, n.d, sent for publication on 1 May 2008; revised form on 18 August 2008.

[24] L. Jäntschi and M.V. Diudea, *Subgraphs of pair vertices*, J. Math. Chem., published online 31 July 2008; DOI 10.1007/5/10910-008-9411-6.

[25] S.D. Bolboacă and L. Jäntschi, *Modelling the property of compounds from structure: statistical methods for models validation*, Environ. Chem. Lett. 6 (2008), pp. 175–181.

[26] T. Colton, *Statistics in Medicine*, Little Brown and Company, New York, NY, 1974.

[27] J.H. Steiger, *Tests for comparing elements of a correlation matrix*, Psychol. Bull. 87 (1980), pp. 245–251.