ET46/2006 – Et. Finală/2006 – Sinteza Lucrării

Ministerul Educaţiei şi Cercetării
Universitatea Tehnică din Cluj-Napoca
Facultatea de Ştiinţa şi Ingineria Materialelor
Catedra de Chimie

Programul:	Cercetare de Excelenţă
Modul: II	Proiecte de Dezvoltare a Resurselor Umane pentru Cercetare
Tipul proiectului:	Proiecte de cercetare de excelenta pentru tinerii cercetători
Cod proiect:	ET46/2006
Denumirea proiectului:	Distribuţia Binomiala: Modelare Statistica, Optimizare Numerica, cu Aplicatii in Bioinformatica si Biochimie
Etapa:	Finală/2006

Lucrare în extenso

Cuprins

Etape, obiective şi activități

Pentru etapele anului 2006 obiectivele planificate au fost:

- *Managementul resurselor de baza ale cercetarii - resurse umane (Etapa 1);*
- *Managementul resurselor de baza ale cercetarii - cunostinte (Etapa 1);*
- *Obtinerea rezultatelor preliminare ale optimizarii, cazul functiilor de o singura variabila binomiala X pentru proportia binomiala simpla X/n (Etapa 2);*
- *Validarea rezultatelor preliminare ale optimizarii, cazul functiilor de o singura variabila binomiala X pentru proportia binomiala simpla X/n (Etapa 2).*

Activitățile prevăzute a se desfăşura au fost:

- *Colectarea de noi metode de calcul al intervalului de confidenta raportate in literatura de specialitate (Etapa 1);*
- *Implementarea si integrarea noilor metode in libraria cu metode deja implementate (Etapa 1)*;
- *Optimizarea vitezei de calcul a algoritmului de calul a erorii experimentale bazat pe distributia binomiala (Etapa 1)*;
- *Reproiectarea modului de calcul al intervalelor de incredere pentru proportii astfel incat sa permita calculul in timp real al intervalelor si pentru n > 500 (Etapa 2)*;
- *Implementarea modului de calcul al intervalelor de incredere pentru proportii astfel incat sa permita calculul in timp real al intervalelor si pentru n > 500 (Etapa 2)*;
- *Optimizarea intervalelor de incredere pentru proportii intervalelor de confidenta pentru n > 500 si fiecare X (Etapa 2);*
- *Participări la manifestări ştiinţifice şi dobândirea de competenţe complementare (Etapele 1 şi 2)*.

Activităţile au fost realizate şi obiectivul planificat a fost atins.

Activităţi şi rezultate

- **Colectarea de noi metode de calcul al intervalului de confidenta raportate in literatura de specialitate (Etapa 1):**

    O nouă metodă de calcul a intervalului de confidenţă pentru populaţii distribuite binomial a fost raportată în literatura de specialitate, metodă care a fost selectată pentru includere în librăria de metode deja existente: *Lawrence BROWN, Xuefeng LI, Confidence intervals for two sample binomial distribution, Journal of Statistical Planning and Inference, 130(1-2), 2005, p. 359-375.*

    De asemenea, a fost studiată lucrarea lui *Miguel A. GARCÍA-PÉREZ, On the Confidence Interval for the Binomial Parameter, Quality and Quantity, 39(4), 2005, p. 467 - 481* care ilustrează prin 3 exemple alegerea corectă a intervalului de confidenţă pentru parametrii binomiali. Şi această lucrare a fost selectată pentru îmbogăţirea librăriei de modele, bazat pe metoda comparativă propusă de aceasta.

- **Implementarea si integrarea noilor metode in libraria cu metode deja implementate (Etapa 1):**

    Au fost implementate 109 funcţii. Următorul listing este obţinut din programul *ci_functions.php* folosind comanda *print_r(get_defined_functions());*:

```
[user] => Array                          [18] => mi
    (                                     [19] => ci
    [0] => ob                             [20] => i
    [1] => betac00                        [21] => r
    [2] => betac10                        [22] => rd
    [3] => betac01                        [23] => wald
    [4] => betac11                        [24] => waldc
    [5] => betacj0                        [25] => wilson
    [6] => betacj1                        [26] => wilsonc
    [7] => betacj2                        [27] => arcs
    [8] => betacja                        [28] => arcsc1
    [9] => betacjb                        [29] => arcsc2
    [10] => betacjc                       [30] => arcsc3
    [11] => betac                         [31] => ac
    [12] => ma2                           [32] => ac1
    [13] => me2                           [33] => acc2
    [14] => me                            [34] => acc3
    [15] => ma                            [35] => logitc
    [16] => mg                            [36] => logit
    [17] => mh                            [37] => bsc
```

[38] => obmin
[39] => obmax
[40] => obmon
[41] => binomialc
[42] => bayes
[43] => cp
[44] => bayesf
[45] => jeffreys
[46] => binomial
[47] => bb00
[48] => bb01
[49] => bb10
[50] => bb11
[51] => bc0
[52] => bc1
[53] => bc2
[54] => bc3
[55] => r2
[56] => r2binomial
[57] => r2binomialc
[58] => r2ac
[59] => r2wald
[60] => r2waldc
[61] => d
[62] => dbinomial
[63] => dbinomialc
[64] => dwald
[65] => dac
[66] => dac1
[67] => dwaldc
[68] => das
[69] => dasc
[70] => ad
[71] => adjeffreys
[72] => adjeffreysc
[73] => adbinomial
[74] => adbinomialc
[75] => adwald
[76] => adac
[77] => adac1
[78] => adwaldc
[79] => adas0
[80] => adas1
[81] => adas2
[82] => iad
[83] => iadwald
[84] => iadac
[85] => iadwaldc
[86] => iadas0
[87] => iadas1

[88] => iadas2
[89] => iadjeffreys
[90] => iadjeffreysc
[91] => iadbinomial
[92] => iadbinomialc
[93] => rp
[94] => rpbin0
[95] => rpwald
[96] => rpac
[97] => rpac1
[98] => rpac0
[99] => rpac2
[100] => arp
[101] => arpw0
[102] => arpac
[103] => waldc1
[104] => arpw1
[105] => waldc2
[106] => arpw2
[107] => waldc3
[108] => arpw3
)

Următoarea secvență de cod implementează funcția BetaC și foloseşte *BetaDistribution.php*:

```
function BetaC($X,$n,$z,$a,$a1,$a2,$c1,$c2){
 if(($X==0)||($X==$n)) {
  if ($X) return array ( pow($a,1/($n)) , 1 );
  if ($n-$X) return array ( 0, 1-pow($a,1/($n)) );
 }
 if($X==0) $Xi = 0; else {
  $binXi = new BetaDistribution($X+$c1,$n-$X+$c2);
  $Xi = $binXi->inverseCDF($a1/2);
 }
 if($X==$n) $Xs = 1; else {
  $binXs = new BetaDistribution($X+$c2,$n-$X+$c1);
  $Xs = $binXs->inverseCDF(1-$a2/2);
 }
 return array ( $Xi , $Xs );
}
```

- **Optimizarea vitezei de calcul a algoritmului de calul a erorii experimentale bazat pe distributia binomiala (Etapa 1);**
- **Reproiectarea modului de calcul al intervalelor de incredere pentru proportii astfel incat sa permita calculul in timp real al intervalelor si pentru n > 500 (Etapa 2);**
- **Implementarea modului de calcul al intervalelor de incredere pentru proportii astfel incat sa permita calculul in timp real al intervalelor si pentru n > 500 (Etapa 2);**
- **Optimizarea intervalelor de incredere pentru proportii intervalelor de confidenta pentru n > 500 si fiecare X (Etapa 2):**

Algoritmul de calcul a erorii experimentale bazat pe distribuția binomială a fost construit folosind librăriile *ProbabilityDistribution.php*, *SpecialMath.php*, *BetaDistribution.php*, *BinomialDistribution.php*.

2 funcții generice calculează eroarea experimentală:

- function est_X($z,$a,$c_i,$f,$g);
- function est_N($z,$a,&$c_i,$cif,$g,$ce);

unde z este percentila de semnificație dată din distribuția Gauss (pentru acele metode care o folosesc), a semnificația dată (nivelul de semnificație α), c_i este funcția de calcul a intervalului de confidență (din cele 109), f respectiv cif este funcția de dependență - $f(x,n) = x/n$, iar g şi ce sunt variabile prevăzute pentru compatibilitate pentru metodele de 2 variabile binomiale (anul de finanțare următor).

Rezultatele obținute în optimizare sunt vizibile online:

http://vl.academicdirect.org/applied_statistics/binomial_distribution/

Rezultatul optimizării pentru n=534 este redat în figurile de mai jos. Figurile sunt reprezentate 1:1 (fiecare pixel reprezintă o valoare a lui X (0..534):
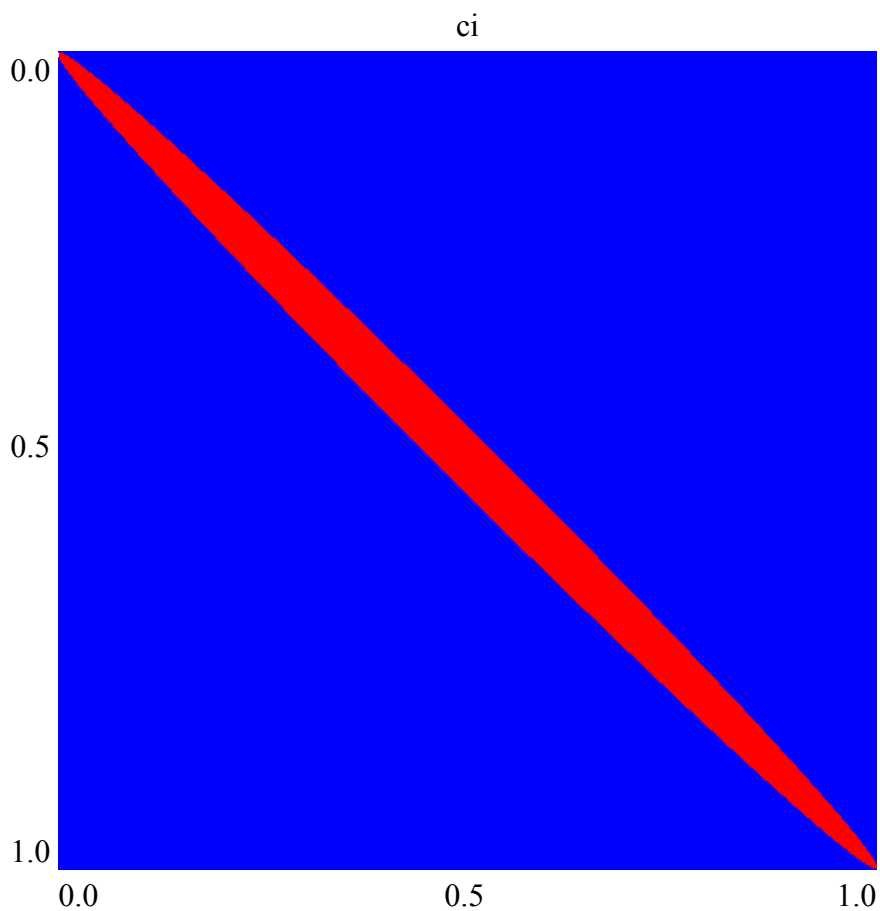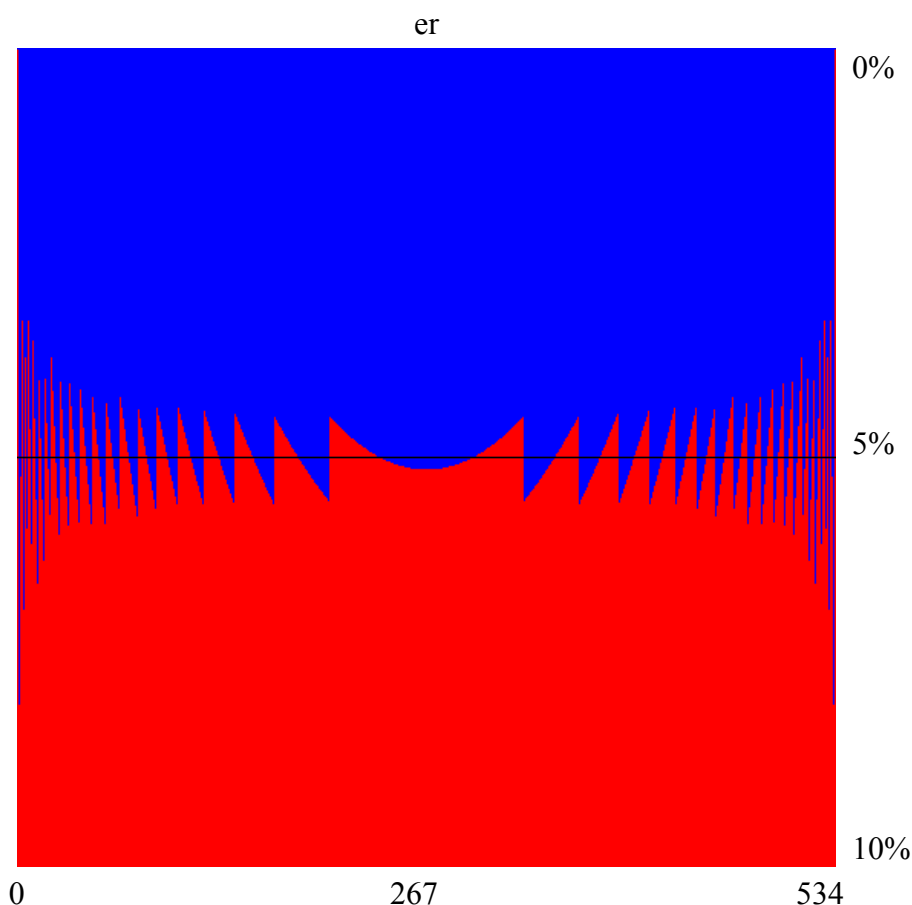
- **Participări la manifestări ştiinţifice şi dobândirea de competenţe complementare (Etapele 1 şi 2):**

S-a participat cu lucrări ştiinţifice la următoarele conferinţe:

1. Carmen Elena STOENOIU, Lorentz JÄNTSCHI, Sorana Daniela BOLBOACĂ, *Computer-Based Testing in Physical Chemistry Topic*, **Third Humboldt Conference on Computational Chemistry**, June 24-28, Conference Proceedings, ISBN 954-323-199-0 then 978-954-323-199-7, p. 94, **Bulgaria, 2006**, Varna. http://lori.academicdirect.org/conferences/3HCCC_Varna_June_2006_2.pdf

2. Sorana BOLBOACĂ, Lorentz JÄNTSCHI, *Computer-Assisted Training and Evaluation System in Evidence-Based Medicine*, **11th International Symposium for Health Information Management Research**, July 14-16, Proceedings, ISBN 0-7703-9016-1, p. 220-226, **Nova Scotia, CA, 2006**, Halifax. http://lori.academicdirect.org/conferences/iSHIMR_220_226.pdf

3. Lorentz JÄNTSCHI, Sorana BOLBOACĂ, *Organizing Guidelines Models and Clinical Practice Guidelines*, **11th International Symposium for Health Information**

Management Research, July 14-16, Proceedings, ISBN 0-7703-9016-1, p. 328-338, **Nova Scotia, CA, 2006**, Halifax.
http://lori.academicdirect.org/conferences/iSHIMR_328_338.pdf

4. Lorentz JÄNTSCHI, Sorana Daniela BOLBOACĂ, *Computer Aided System for Student's Knowledge Assessment*, **The 10th World Multi-Conference on Systemics, Cybernetics and Informatics**, July 16-19, Proceedings, ISBN 980-6560-65-5 (Collection) && ISBN 980-6560-65-3 (Volume 1), p. 97-101, **Florida, U.S.A, 2006**, Orlando.
http://lori.academicdirect.org/conferences/Wmsci2006_1.pdf

5. Sorana Daniela BOLBOACĂ, Lorentz JÄNTSCHI, *Assessment of an Computer Based Curriculum in Evidence-Based Medicine*, **The 10th World Multi-Conference on Systemics, Cybernetics and Informatics**, July 16-19, e-Proceedings, ISBN 980-6560-92-2 (CD), paper #3 e-KCC (THEME: Other theme or topic in the domain of KCC 2006), 5 p., **Florida, U.S.A, 2006**, Orlando.
http://lori.academicdirect.org/conferences/Wmsci2006_2.pdf

6. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Interactive Web Application for Evidence-Based Medicine Training*, **MEDNET 2006 Abstracts Book (11th World Congress on Internet in Medicine)**, October 14-19, abstract published in volume at page 63, **Canada, 2006**, Toronto. http://lori.academicdirect.org/conferences/draft-mednet2006book.pdf
http://lori.academicdirect.org/conferences/MedNet06_Jantschi&Bolboaca.pdf

7. Sorana-Daniela BOLBOACĂ, ACHIMAŞ CADARIU, Ştefan ŢIGAN, *Distance Learning Strategy: Continuing Medical Education on Evidence-Based Medicine - Implementation at Cluj County*, **11th World Congress on Internet in Medicine**, October 14-19, Toronto, Ontario, **Canada, 2006**, abstract published in volume at page 24.
http://sorana.academicdirect.ro/conferences/draft-mednet2006book.pdf
http://sorana.academicdirect.ro/conferences/MedNet06_Bolboaca&All.pdf

8. Sorana BOLBOACĂ, Lorentz JÄNTSCHI, Ştefan ŢIGAN, *Communication of Results on Risk Factors Studies: Confidence Intervals*, **ISCB27 - International Society for Clinical Biostatistics**, 2006 August 27-31, Geneva, Switzerland, Conference Program, abstract no. 273. http://lori.academicdirect.org/conferences/Geneva06.pdf
http://lori.academicdirect.org/conferences/ISCB_abs273-Bolboaca.pdf

9. Lorentz JÄNTSCHI, Carmen Elena STOENOIU, Sorana Daniela BOLBOACĂ, *Undergraduate Students Assessment on Materials Chemistry Topic using an Auto-Calibrated Online System*, **International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CIS²E 06)**, Online, December, 2006. http://lori.academicdirect.org/conferences/EIAE06_Jantschi&All.pdf

Publicații

1. Sorana BOLBOACĂ, Lorentz JÄNTSCHI, *Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds*, **Leonardo Journal of Sciences**, **AcademicDirect**, Internet, Issue 9, 179-200, 2006, ISSN 1583-0233, DOAJ, http://ljs.academicdirect.ro/A09/179_200.pdf

2. Lorentz JÄNTSCHI, Sorana BOLBOACĂ, *Auto-calibration Online Evaluation: Database Design and Implementation*, **Leonardo Electronic Journal of Practices and Technologies**, **AcademicDirect**, Internet, Issue 9, 179-192, 2006, ISSN 1583-1078, DOAJ, http://lejpt.academicdirect.ro/A09/179_192.pdf

3. Lorentz JÄNTSCHI, Mircea V. DIUDEA, *Subgraphs of Pair Vertices*, **Journal of Mathematical Chemistry**, trimisă spre publicare. http://lori.academicdirect.org/articles/Subgraphs_Pair_Vertices.pdf

4. *Sorana-Daniela BOLBOACĂ, Lorentz JÄNTSCHI, Computer-Based Testing on Physical Chemistry Topic: A Case Study*, **International Journal of Education and Development using Information and Communication Technology (IJEDICT)***, trimisă spre publicare. http://lori.academicdirect.org/articles/IJEDITC_Bolboaca&Jantschi.pdf*

5. *Lorentz JÄNTSCHI, Sorana Daniela BOLBOACĂ, Communication on Polynomial Formula for Vertex Cuts in B-Trees*, **Kragujevac Journal of Mathematics***, trimisă spre publicare. http://lori.academicdirect.org/articles/KragujevacJournalOfMathematics.pdf*

6. *Sorana-Daniela BOLBOACĂ, Lorentz JÄNTSCHI, Ştefan ŢIGAN, Communication of Results on Risk Factors Studies: Confidence Intervals*, **Applied Medical Informatics***, trimisă spre publicare. http://lori.academicdirect.org/articles/CommResultsOnRiskFactorsStudies.pdf*

7. Lorentz JÄNTSCHI, Mădălina Ana VĂLEANU, Sorana Daniela BOLBOACĂ, *Programarea rapida a aplicatiilor pentru baze de date relationale (in Romanian)*, AcademicDirect & AcademicPres, **Internet & Cluj-Napoca**, 233 p., ISBN(10) 973-86211-5-1 & ISBN(13) 978-973-86211-5-2 (AcademicDirect) && ISBN(10) 973-744-044-7 & ISBN(13) 978-973-744-044-0 (AcademicPres), *2006 (November)* *http://ph.academicdirect.org/*

Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds

Abstract

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was studied by the use of molecular descriptors family on structure-activity relationships. Starting from the results obtained by applying of MDF-SAR methodology on pyrimidine derivatives and from the assumption that the measured activity (compounds' inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used, etc.), the abilities of Pearson, Spearman, Kendall's, and Gamma correlation coefficients in analysis of estimated toxicity were studied and are presented.

Keywords

Multiple linear regressions, Correlation coefficients, Molecular Descriptors Family on Structure-Activity Relationships (MDF-SAR)

Introduction

QSAR (**Q**uantitative **S**tructure-**A**ctivity **R**elationships) is an approach which is able to indicate for a given compound or a class of compounds which feature of structure characteristics is correlated with its activity [1]. In QSAR analysis were proposed several approaches for development. Simple and multiple linear regressions is one of the more successful techniques use by many researcher in construct of QSAR models [2-4].

[1] Rogers D., Hopfinger A. J., *Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships*, J. Chem. Inf. Comput. Sci. 34, 1994, p. 854-866.

[2] Hansch C., Leo A., Stephen R., Eds. Heller, *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*, ACS professional Reference Book., American

Correlation coefficient is a simple statistical measure of relationship between one dependent and one or more than one independent variables and it is use as a measure of the statistical fit of a regression based model in QSAR [5]. Its squared value (the coefficient of determination) it is most frequently used parameter as a measure of the goodness-of-fit of the model [6-10].

Chemical Society, Washington, D.C., 1995.

[3] Zahouily M., Lazar M., Elmakssoudi A., Rakik J., Elaychi S., Rayadh A., *QSAR for anti-malarial activity of 2-aziridinyl and 2,3-bis(aziridinyl)-1,4-naphthoquinonyl sulfonate and acylate derivatives*, J Mol Model 12(4), 2006, p. 398-405.

[4] Liang G.-Z., Mei H., Zhou P., Zhou Y., Li Z.-L., *Study on quantitative structure-activity relationship by 3D holographic vector of atomic interaction field*, Acta Phys-Chim Sin 22(3), 2006, p. 388-390.

[5] Rosner B., *Fundamentals of Biostatistics*, 4th Edition, Duxbury Press, Belmont, California, USA, 1995.

[6] Katritzky A. R., Kuanar M., Slavov S., Dobchev D.A., Fara D. C., Karelson M., Acree Jr. W. E., Solov'ev V. P., Varnek A., *Correlation of blood-brain penetration using structural descriptors*, Bioorg Med Chem, 14(14), 2006, p. 4888-4917.

[7] Wang Y., Zhao C., Ma W., Liu H., Wang T., Jiang G., *Quantitative structure-activity relationship for prediction of the toxicity of polybrominated diphenyl ether (PBDE) congeners*, Chemosphere 64(4), 2006, p. 515-524.

[8] Roy D. R., Parthasarathi R., Subramanian V., Chattaraj P. K., *An electrophilicity based analysis of toxicity of aromatic compounds towards Tetrahymena pyriformis*, QSAR Comb Sci 25(2), 2006, p 114-122.

[9] Srivastava H. K., Pasha F. A., Singh P. P., *Atomic softness-based QSAR study of testosterone*, Int J Quantum Chem 103(3), 2005, p. 237-245.

[10] Xue C. X., Zhang R. S., Liu H. X., Yao X. J., Hu M. C., Hu Z. D., Fan B. T., *QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine*, J Chem Inf Comput Sci 44(5), 2004, p. 1693-1700.

A new approach of molecular descriptors family on structure-activity relationships (MDF-SAR) was developed [11], and proved its usefulness in estimation and prediction of: toxycity [12, 13], mutagenicity [12], antioxidant efficacy [14], antituberculotic activity [15], antimalarial activity [16], antiallergic activity [17], anti-HIV-1 potencies [18], inhibition activity on carbonic anhydrase II [19] and IV [20].

[11] Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology*, Leonardo Electronic Journal of Practices and Technologies 6, 2005, p. 76-98.

[12] Jäntschi L., Bolboacă S., *Molecular Descriptors Family on QSAR Modeling of Quinoline-based Compounds Biological Activities*, The 10th Electronic Computational Chemistry Conference 2005; http://bluehawk.monmouth.edu/~rtopper/eccc10_absbook.pdf as on 13 May 2006.

[13] Bobloacă S.D., Jäntschi L., *Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information*, Terapeutics, Pharmacology and Clinical Toxicology X(1), 2006, p. 110-114.

[14] Bolboacă S., Filip C., Ţigan Ş., Jäntschi L., *Antioxidant Efficacy of 3-Indolyl Derivates by Complex Information Integration*, Clujul Medical LXXIX(2), 2006, p. 204-209.

[15] Bolboacă S., Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 3. Antituberculotic Activity of some Polyhydroxyxanthones*, Leonardo Journal of Sciences 7, 2005, p. 58-64.

[16] Jäntschi L., Bolboacă S., *Molecular Descriptors Family on Structure Activity Relationships 5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates*, Leonardo Journal of Sciences 8, 2006, p. 77-88.

[17] Jäntschi L., Bolboacă S., Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction, Clujul Medical LXXIX, 2006, In press.

[18] Bolboacă S., Ţigan Ş., Jäntschi L., *Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives*, In: Reichert A., Mihalaş G., Stoicu-Tivadar L., Schulz Ş., Engelbrech R. (Eds.), Proceedings of the European Federation for Medical Informatics Special Topic Conference, p. 222-226, 2006.

[19] Jäntschi L., Ungureşan M. L., Bolboacă S.D., *Integration of Complex Structural Information in Modeling of Inhibition Activity on Carbonic Anhydrase II of Substituted Disulfonamides*, Applied Medical Informatics 17(3,4), 2005, p. 12-21.

Several correlation coefficients based on different statistical hypothesis are known and most frequently used today: Pearson correlation coefficient, Spearman rank correlation coefficient and Spearman semi-quantitative correlation coefficient, Kendall tau-a, -b and -c correlation coefficients, Gamma correlation coefficient [5].

Starting from the results obtained by applying of MDF-SAR methodology on a sample of sixty-seven compounds and from the assumption that the measured activity (compounds' inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used), the abilities of Pearson, Spearman, Kendall's, and Gamma correlation coefficients in analysis of estimated toxicity were studied.

Multi-varied MDF-SAR model of pyrimidine derivatives

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was studied by the use of MDF-SAR methodology.

The set of pyrimidine derivatives (2,4-Diamino-5-(substituted-benzyl)-pyrimidine derivatives) with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was previously studied by Ting-Lan Chiu & Sung-Sau So by the use of neural network approach [21].

By applying the MDF-SAR methodology on the sample of sixty-seven pyrimidine derivatives, a multi-varied model with four descriptors reveled to has good performances in prediction and estimation of inhibitory activity.

The multi-varied MDF-SAR model with four descriptors had the following equation:

$$Y_{est} = 3.78 + 1.62 \cdot iImrKHt + 2.37 \cdot liMDWHg + 6.40 \cdot IsDrJQt - 8.52 \cdot 10^{-2} \cdot LSPmEQg$$

Analyzing the MDF-SAR model with four descriptors it could be say that inhibitory activity consider compounds geometry (**g**) and topology (**t**), being related with the number of

[20] Jäntschi L., Bolboacă S., *Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information*, Electronic Journal of Biomedicine, 2006, In press.

[21] Chiu T.L., So S. S., *Development of neural network QSPR models for Hansch substituent constants. 2. Applications in QSAR studies of HIV-1 reverse transcriptase and dihydrofolate reductase inhibitors*, J Chem Inf Comput Sci 44(1), 2004, p. 154-160.

directly bonded hydrogen's (**H)** of compounds and with the partial charge (**Q**) as atomic properties.

Statistical characteristics of the MDF-SAR model with four descriptors are in table 1 and 2.

Table 1. Statistical characteristics of the multi-varied MDF-SAR model with four descriptors

Characteristic (notation)	Value
Number of variable (v)	4
Correlation coefficient (r)	0.9517
95% Confidence Intervals for r (95% CI$_r$)	[0.9223, 0.9701]
Squared correlation coefficient (r^2)	0.9058
Adjusted squared correlation coefficient (r$^2_{adj}$)	0.8997
Standard error of estimated (s$_{est}$)	0.1919
Fisher parameter (F$_{est}$)	149[*]
Cross-validation leave-one-out (loo) score (r$^2_{cv-loo}$)	0.8932
Fisher parameter for loo analysis (F$_{pred}$)	130[*]
Standard error for leave-one-out analysis (s$_{loo}$)	0.2044
Model stability (r^2 - r$^2_{cv(loo)}$)	0.0126
r^2(*iImrKHt, liMDWHg*)	0.2020
r^2(*iImrKHt, IsDrJQt*)	0.0047
r^2(*iImrKHt, LSPmEQg*)	0.1482
r^2(*liMDWHg, IsDrJQt*)	0.0003
r^2(*liMDWHg, LSPmEQg*)	0.0212
r^2(*IsDrJQt, LSPmEQg*)	0.0664

[*]p < 0.001

Table 2. Statistics of the regression MDF-SAR model with four descriptors

	StdError	t Stat	95%CI$_{coefficient}$	r(Y$_m$,desc)
Intercept	0.1999	18.92[*]	[3.38, 4.18]	n.a.
iImrKHt	0.0709	22.85[*]	[1.48, 1.76]	0.4803
liMDWHg	0.1500	15.81[*]	[2.07, 2.67]	0.0558
IsDrJQt	1.4779	4.33[*]	[3.45, 9.36]	0.0336
LSPmEQg	0.0182	-4.68[*]	[-0.12, -0.12]	0.0231

StdError = standard error; t Stat = Student tets parameter;
95% CI$_{coefficient}$ = 95% confidence interval associated with regression coefficients;
Y$_m$ = measured inhibitory activity; desc = molecular descriptor; [*] p < 0.001

Graphical representation of the measured versus estimated by MDF-SAR model with four descriptors inhibitory activity is in figure 1.
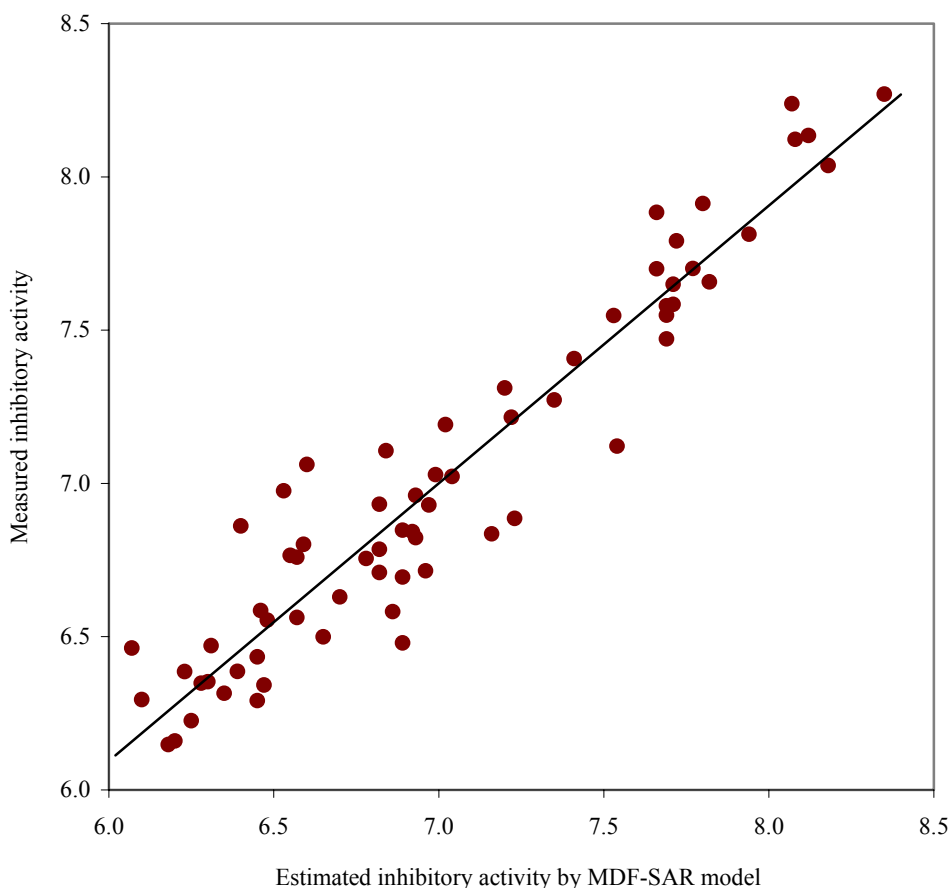
Figure 1. Plot of measured vs estimated by MDF-SAR inhibitory activity

Internal validation of the four-varied MDF SAR model with four descriptors was performed through splitting the whole set into training and test sets by applying of a randomization algorithm.

The coefficients for each model obtained in training sets, in conformity with the generic equation $Y_{est} = a_0 + a_1 \cdot iImrKHt + a_2 \cdot liMDWHg + a_3 \cdot IsDrJQt - a_4 \cdot 10^{-2} \cdot LSPmEQg$, the number of compounds in training (N_{tr}) and test (N_{ts}) sets, the correlation coefficient for training (r_{tr}) and test (r_{ts}) sets with associated 95% confidence intervals (95%CI_{rtr} and 95%CI_{rts}), the Fisher parameter associated with training (F_{tr}) and test (F_{ts}) sets, and the Fisher's Z parameter of correlation coefficients comparison ($Z_{rtr-rts}$) are in table 3.

Table 3. Statistics results on training versus test sets

a_0	a_1	a_2	a_3	a_4	N_{tr}	r_{tr}	95%CI_{rtr}	F_{tr}	N_{ts}	r_{ts}	95%CI_{ts}	F_{ts}	$Z_{rtr-rts}$
3.93	1.61	2.43	6.56	$-9.67 \cdot 10^{-2}$	35	0.949	[0.899, 0.974]	67[*]	32	0.958	[0.916, 0.980]	59[*]	0.418[†]
3.98	1.57	2.45	6.55	$-7.52 \cdot 10^{-2}$	36	0.951	[0.905, 0.975]	73[*]	31	0.951	[0.899, 0.976]	61[*]	0.000[†]
3.84	1.55	2.15	9.08	$-9.12 \cdot 10^{-2}$	37	0.944	[0.893, 0.908]	66[*]	30	0.949	[0.895, 0.976]	55[*]	0.206[†]
3.94	1.59	2.42	6.10	$-8.18 \cdot 10^{-2}$	38	0.951	[0.907, 0.974]	78[*]	29	0.947	[0.890, 0.975]	50[*]	0.144[†]
3.91	1.56	2.25	8.22	$-1.04 \cdot 10^{-1}$	39	0.963	[0.931, 0.981]	110[*]	28	0.937	[0.867, 0.971]	39[*]	1.069[†]

4.18	1.51	2.44	6.06	$-7.22 \cdot 10^{-2}$	40	0.956	[0.917, 0.975]	92*	27	0.936	[0.863, 0.971]	35*	0.721†
3.76	1.63	2.32	7.35	$-1.02 \cdot 10^{-1}$	41	0.963	[0.931, 0.980]	116*	26	0.935	[0.858, 0.971]	34*	1.104†
3.97	1.58	2.39	5.11	$-9.36 \cdot 10^{-2}$	42	0.956	[0.919, 0.976]	99*	25	0.954	[0.896, 0.980]	34*	0.115†
3.64	1.64	2.30	7.00	$-8.15 \cdot 10^{-2}$	43	0.955	[0.917, 0.975]	98*	24	0.944	[0.873, 0.976]	37*	0.407†
3.72	1.66	2.43	5.78	$-8.12 \cdot 10^{-2}$	44	0.938	[0.889, 0.966]	72*	23	0.964	[0.916, 0.985]	54*	1.030†
3.59	1.64	2.25	4.94	$-9.98 \cdot 10^{-2}$	45	0.947	[0.904, 0.970]	86*	22	0.957	[0.898, 0.982]	37*	0.411†
3.86	1.55	2.23	8.68	$-8.86 \cdot 10^{-2}$	46	0.940	[0.894 0.967]	78*	21	0.983	[0.958, 0.993]	43*	**2.290***
4.04	1.54	2.36	6.46	$-7.31 \cdot 10^{-2}$	47	0.949	[0.911, 0.972]	96*	20	0.963	[0.906, 0.985]	34*	0.538†
3.63	1.63	2.24	4.27	$-8.93 \cdot 10^{-2}$	48	0.940	[0.895, 0.966]	82*	19	0.963	[0.904, 0.986]	44*	0.852†
3.98	1.57	2.42	6.49	$-8.59 \cdot 10^{-2}$	49	0.946	[0.905, 0.969]	93*	18	0.960	[0.894, 0.985]	36*	0.535†
3.77	1.61	2.32	6.37	$-8.46 \cdot 10^{-2}$	50	0.943	[0.902, 0.968]	91*	17	0.974	[0.927, 0.991]	52*	1.294†
3.67	1.63	2.22	6.56	$-1.01 \cdot 10^{-1}$	51	0.954	[0.919, 0.973]	115*	16	0.950	[0.858, 0.983]	17*	0.126†
3.81	1.61	2.39	6.87	$-7.70 \cdot 10^{-2}$	52	0.951	[0.916, 0.972]	112*	15	0.950	[0.853, 0.984]	22*	0.032†
3.69	1.65	2.36	6.32	$-8.21 \cdot 10^{-2}$	53	0.953	[0.919, 0.972]	118*	14	0.956	[0.864, 0.986]	17*	0.128†
3.97	1.56	2.40	6.16	$-7.51 \cdot 10^{-2}$	54	0.951	[0.916, 0.971]	115*	13	0.954	[0.851, 0.987]	17*	0.122†

† $p > 0.05$; * $p < 0.01$

Definitions, Formulas, Interpretations, PHP functions, and Results

A number of add notations were used in the study, as follows:

- Pearson product-moment correlation coefficient (named after Karl Pearson (1857 - 1936), a major contributor to the early development of statistics):
  - r_{prs} = the *Pearson* correlation coefficient;
  - r_{Prs}^2 = the squared *Pearson* correlation coefficient;
  - $t_{Prs,df}$ = the Student test parameter, and its significance $p_{Prs,df}$ at a significance level of 5% (where df = the degree of freedom);
- Spearman's rank correlation coefficient (named after Charles Spearman (1863 - 1945), English psychologist known for his work in statistics - *factor analysis*, and *Spearman's rank correlation coefficient*):
  - r_{Spm} = the *Spearman* rank correlation coefficient
  - r_{Spm}^2 = the squared of *Spearman* rank correlation coefficient;
  - $t_{Prs,df}$ = the Student test parameter, and its significance $p_{Spm,df}$;
  - r_{sQ}^2 = the squared of *Spearman semi-Qantitative* correlation coefficient;
  - t_{sQ} = the Student test parameter, and its significance p_{sQ};
- Kendall's tau correlation coefficients (named after Maurice George Kendall (1907 - 1983), a prominent British statistician; published in monograph *Rank Correlation* in 1948);
  - $\tau_{Ken,a}$ = the *Kendall tau-a* correlation coefficient;
  - $\tau_{Ken,a}^2$ = the squared of *Kendall tau-a* correlation coefficient;

- o $Z_{Ken,\tau a}$ = the Z-test parameter of Kendall tau-a correlation coefficient, and its significance $p_{Ken,\tau a}$;
- o $\tau_{Ken,b}$ = the *Kendall tau-b* correlation coefficient;
- o $\tau_{Ken,b}^2$ = the squared of *Kendall tau-b* correlation coefficient;
- o $Z_{Ken,\tau b}$ = the Z-test parameter of Kendall tau-b, and its significance $p_{Ken,\tau b}$;
- o $\tau_{Ken,c}^2$ = the *Kendall tau-c* correlation coefficient;
- o $\tau_{Ken,c}^2$ = the squared of *Kendall tau-c* correlation coefficient;
- o $Z_{Ken,\tau c}$ = the Z-test parameter of *Kendall tau-c*, and its significance $p_{Ken,\tau c}$;

- Gamma correlation coefficient (also known as Goodman and Kruskal's gamma):
  - o Γ = the *Gamma* correlation coefficient;
  - o Γ^2 = the squared of *Gamma* correlation coefficient;
  - o Z_Γ = the Z-test parameter of *Gamma* correlation coefficient, and its significance p_Γ.

A series of **.php* programs which to facilitate the calculation and to display of above-described correlation coefficients and their statistics (Student-test and Z-test parameters and associated significances) were implemented and was use in order to reach the objective of study [22].

Pearson correlation coefficient

Definition: a measure the strength and direction of the linear relationship between two variables, describing the direction and degree to which one variable is linearly related to another.

Assumptions: both variable (variables Y_m and Y_{est}) are interval or ratio variables and are well approximated by a normal distribution, and their joint distribution is bivariate normal [23].

[22] ***Rank, ©2005, Virtual Library of Free Software, available at: http://vl.academicdirect.org/molecular_topology/mdf_findings/rank/

[23] ***Pearson's Correlation Coefficient [online], Available at: http://www.texasoft.com/winkpear.html

Formula

$$r_{Prs} = \frac{\sum (Y_{m-i} - \overline{Y}_m)(Y_{est-i} - \overline{Y}_{est})}{\sqrt{\left(\sum (Y_{m-i} - \overline{Y}_m)^2\right)\left(\sum (Y_{est-i} - \overline{Y}_{est})^2\right)}}$$

where Y_{m-i} is the value of the measured inhibitory activity for compound i (i = 1, 2, …, 67) $\overline{Y}_m$ is the average of the measured inhibitory activity, Y_{est-i} is the value of the estimated inhibitory activity for compound i, and $\overline{Y}_{est}$ is the average of the estimated inhibitory activity.

Interpretation

The Pearson correlation coefficient can take values from -1 to +1. A value of +1 show that the variables are perfectly linear related by an increasing relationship, a value of -1 show that the variables are perfectly linear related by an decreasing relationship, and a value of 0 show that the variables are not linear related by each other. There is considered a strong correlation if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5.

The coefficient of determination (or r squared) gives information about the proportion of variation in the dependent variable which might be considered as being associated with the variation in the independent variable.

Related statistics

- The squared of Pearson correlation coefficient or Pearson coefficient of determination (r_{Prs}^2);
  o Describe the proportion of variance in Y_m that is related with linear variation of Y_{est};
  o Can take values from 0 to 1.

Statistical test

Student t-test was used to determine if the value of Pearson correlation coefficient is statistically significant, at a significance level of 5%.

The null hypothesis vs. the alternative hypothesis was:

H_0: $r_{Prs} = 0$ (there is no correlation between the variables)

H_1: $r_{Prs} <> 0$ (variables are correlated)

For a significance level equal with 5%, a p-value associated to $t_{Prs,df}$ less than 0.05 means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis. In other words there is a statistically significant linear relationship between the variables.

PHP implementation

In order to compute the statistics associated with Pearson correlation coefficient, three functions were implemented:

```php
function coef_rk(&$y1,&$y2){
  $my1=m1($y1);
  $dy2=m2($y1,$y1)-$my1*$my1;
  $mx1=m1($y2);
  $mxy=m2($y2,$y1);
  $m2x=$mx1*$mx1;
  $mx2=m2($y2,$y2);
  $dx2=$mx2-$m2x;
  $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
  return $r2;
}
function t_p($n,$k,$r){
  return $r*pow($n-$k-1,0.5)/pow(1-pow($r,2),0.5);
}
function p_t($t,$df){
  $p = $df/2;
  $x = 0.5+0.5*$t/pow(pow($t,2)+$df,0.5);
  $beta_gam = exp( -logBeta($p, $p) + $p * log($x) + $p * log(1.0 - $x) );
  return (2.0 * $beta_gam * betaFraction(1.0 - $x, $p, $p) / $p);
}
```

The statistics of Pearson correlation coefficients are computed as follows:

- Pearson correlation coefficient:

    $r_pe = coef_rk($cmp[0],$cmp[1]);

where *$cmp[0]* is the measured inhibitory activity (Y_m), and *$cmp[1]* is the estimated by MDF-SAR model with four descriptor inhibitory activity (Y_{est}).

- t Student parameter:

    $t_pe = t_p($n,1,pow($r_pe,0.5));

- Significance of t Student parameter

    $p_pe = p_t($t_pe,$n-2);

    ### Results

    $r_{Prs}^2 = 0.9058$

$$t_{Prs,1} = 24.99 \tag{1}$$

$$p_{Prs,1} = 4.74 \cdot 10^{-33} \, \%$$

Spearman's rank correlation coefficient

Definition

A non-parametric measure of correlation between variable which assess how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Frequently the Greek letter ρ (rho) is use to abbreviate the Spearman correlation coefficient.

Spearman's rank correlation is satisfactory for testing the null hypothesis of no relationship, but is difficult to interpret as a measure of the strength of the relationship [24].

Assumptions

- Does not required any assumptions about the frequency distribution of the variables;
- Does not required the assumption that the relationship between variable is linear;
- Does not required the variable to be measured on interval or ration scale.

Formula

In order to compute the Spearman rank correlation coefficient, the two variables (Y_m, respectively Y_{est}) were converted to ranks (see table 4 for exemplification). For each measured and estimated inhibitory activity a rank was assigned ($RankY_m$ - for measure inhibitory activity, $RankY_{est}$ - for estimated by MDF-SAR model inhibitory activity) according with the position of value into a sort serried of values.

In assignment of rank process, the lowest value had the lowest rank. When there are two equal values for two different compounds (for measured and/or estimated inhibitory activity), the associated rank had equal values and was calculated as means of corresponding ranks. For example, the compounds abbreviated as c_52 and c_59 have the same measured inhibitory activity (6.45, see table 4). The rank associated with these values is equal with 13.5 (is the average between the rank for c_52 - 13 and the rank of c_59 - 14).

---

[24] Methods based on rank order. In: Bland M., *An Introduction to Medical Statistics*, Oxford University Press; Oxford, New York, Tokyo, p. 205-225, 1995.

ET46/2006 – Et. Finală/2006 – Lucrare in extenso

Table 4. Compounds abbreviation, measured and estimated activity and associated ranks

Abb.	Y_m	RankY_m		Y_{est}	RankY_{est}	Abb.	Y_m	RankY_m		Y_{est}	RankY_{est}
c_64	6.07	1	0	6.4626	13	c_32	6.92	35	0	6.8423	32
c_65	6.10	2	0	6.2948	5	c_66	6.93	36.5	5	6.8225	30
c_67	6.18	3	0	6.1479	1	c_36	6.93	36.5		6.9609	38
c_54	6.20	4	0	6.1595	2	c_40	6.96	38	0	6.7150	24
c_37	6.23	5	0	6.3859	10	c_17	6.97	39	0	6.9298	36
c_48	6.25	6	0	6.2254	3	c_45	6.99	40	0	7.0283	41
c_31	6.28	7	0	6.3483	8	c_41	7.02	41	0	7.1919	45
c_49	6.30	8	0	6.3528	9	c_15	7.04	42	0	7.0225	40
c_10	6.31	9	0	6.4703	14	c_28	7.16	43	0	6.8355	31
c_56	6.35	10	0	6.3149	6	c_09	7.20	44	0	7.3115	48
c_47	6.39	11	0	6.3866	11	c_18	7.22	45	0	7.2156	46
c_53	6.40	12	0	6.8614	34	c_43	7.23	46	0	6.8855	35
c_52	6.45	13.5	1	6.2913	4	c_29	7.35	47	0	7.2724	47
c_59	6.45	13.5		6.4336	12	c_14	7.41	48	0	7.4072	49
c_16	6.46	15	0	6.5851	20	c_24	7.53	49	0	7.5476	51
c_34	6.47	16	0	6.3422	7	c_22	7.54	50	0	7.1218	44
c_58	6.48	17	0	6.5536	17	c_26	7.66	51.5	6	7.7002	57
c_35	6.53	18	0	6.9755	39	c_08	7.66	51.5		7.8841	61
c_42	6.55	19	0	6.7654	27	c_27	7.69	54	7	7.4715	50
c_30	6.57	20.5	2	6.5625	18	c_13	7.69	54		7.5489	52
c_61	6.57	20.5		6.7594	26	c_12	7.69	54		7.5793	53
c_33	6.59	22	0	6.8010	29	c_04	7.71	56.5	8	7.5841	54
c_51	6.60	23	0	7.0616	42	c_11	7.71	56.5		7.6497	55
c_39	6.65	24	0	6.4993	16	c_19	7.72	58	0	7.7915	59
c_38	6.70	25	0	6.6297	21	c_23	7.77	59	0	7.7014	58
c_57	6.78	26	0	6.7552	25	c_25	7.80	60	0	7.9130	62
c_60	6.82	28	3	6.7091	23	c_01	7.82	61	0	7.6576	56
c_44	6.82	28		6.7847	28	c_21	7.94	62	0	7.8130	60
c_55	6.82	28		6.9318	37	c_06	8.07	63	0	8.2391	66
c_20	6.84	30	0	7.1067	43	c_03	8.08	64	0	8.1224	64
c_46	6.86	31	0	6.5813	19	c_07	8.12	65	0	8.1353	65
c_50	6.89	33	4	6.4794	15	c_05	8.18	66	0	8.0372	63
c_62	6.89	33		6.6942	22	c_02	8.35	67	0	8.2702	67
c_63	6.89	33		6.8475	33						

The method of rank assignment for more then two equal values of measured and/or estimated inhibitory activity is the same as for two equal values. If there are an odd number of compounds which have the same measured value (see compounds c_60, c_44, and c_55 from table 2) then the rank will be an integer ((27+28+29)/3 = 28, see the rank for c_60, c_44, and c_55).

In studied example, there are equal values for measured activity: five situations of two equal values (c_52-c_59, c_30-c_61, c_66-c_36, c_26-c_08, and c_04-c_11), and three situations of three equal values (c_60-c_44-c_55, c_50-c_62-c_63, and c_27-c_13-c_12).

By conversion of the measured and estimated inhibitory activity to ranks, the distribution of ranks does not depend on the distribution of measured, respectively estimated inhibitory activity.

The formula for calculation of the Spearman rank correlation coefficient is:

$$r_{Spm} = \frac{\sum (R_{Y_{m-i}} - \overline{R}_{Y_m})(R_{Y_{est-i}} - \overline{R}_{Y_{est}})}{\sqrt{\left(\sum (R_{Y_{m-i}} - \overline{R}_{Y_m})^2\right)\left(\sum (R_{Y_{est-i}} - \overline{R}_{Y_{est}})^2\right)}}$$

where R_{Ym-i} is the rank of the measured inhibitory activity for compound i, $\overline{R}_{Y_{m-i}}$ is the average of the measured inhibitory activity, R_{Yest-i} is the rank of the estimated by MDF-SAR inhibitory activity for compound i, and $\overline{R}_{Y_{est-i}}$ is the average of the estimated inhibitory activity.

The simple formula for r_{Spm} is based on the difference between each pairs of ranks:

$$r_{Spm} = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

where D is the differences between each pair of ranks (e.g. $D = R_{Ym-1} - R_{Yest-1}$) and n is the volume of the sample.

The formula of the *Spearman semi-quantitative method* is:

$$r_{sQ} = \sqrt{\frac{\sum (Y_{m-i} - \overline{Y}_m)(Y_{est-i} - \overline{Y}_{est})}{\sqrt{\left(\sum (Y_{m-i} - \overline{Y}_m)^2\right)\left(\sum (Y_{est-i} - \overline{Y}_{est})^2\right)}} \cdot \frac{\sum (R_{Y_{m-i}} - \overline{R}_{Y_m})(R_{Y_{est-i}} - \overline{R}_{Y_{est}})}{\sqrt{\left(\sum (R_{Y_{m-i}} - \overline{R}_{Y_m})^2\right)\left(\sum (R_{Y_{est-i}} - \overline{R}_{Y_{est}})^2\right)}}}$$

Interpretation

- Identical with Pearson correlation coefficient.

Related statistics

- r_{Spm}^2 = the squared of *Spearman* rank correlation coefficient;
- r_{sQ}^2 = the squared of semi-quantitative correlation coefficient.

Statistical significance

- Compute by the use of a permutation test (a statistical test in which the reference distribution is obtained by permuting the observed data points across all possible outcomes, given a set of conditions consistent with the null hypothesis);

- Comparing the observed r_{Spm} with published tables for different levels of significance (eg. 0.05, 0.01…). It is a simple solution when the researchers want to know the significance within a certain range or less than a certain value;

- Tested by applying the Student t-test (for sample sizes > 20): the method used in this study.

The null hypothesis vs. the alternative hypothesis for Spearman rank correlation coefficient was:

H_0: $r_{Spm} = 0$ (there is no correlation between the ranked pairs)

H_1: $r_{Spm} <> 0$ (ranked pairs are correlated)

The null hypothesis vs. the alternative hypothesis for semi-quantitative correlation coefficient was:

H_0: $r_{sQ} = 0$ (there is no correlation between the ranked pairs)

H_1: $r_{sQ} <> 0$ (ranked pairs are correlated)

PHP implementation

The formulas for Spearman and respectively semi-quantitative correlation coefficients used two defined above functions (*t_p* and respectively *p_t*). The Spearman rank correlation coefficient used the *coef_rk* function defined as:

```
function coef_rk(&$y1,&$y2){
  $my1=m1($y1);
  $dy2=m2($y1,$y1)-$my1*$my1;
  $mx1=m1($y2);
  $mxy=m2($y2,$y1);
  $m2x=$mx1*$mx1;
  $mx2=m2($y2,$y2);
  $dx2=$mx2-$m2x;
  $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
  return $r2;
}
```

where

```
function m1(&$v){
  $rez=0;
  $n=count($v);
  for($i=1;$i<$n;$i++)
    $rez+=$v[$i];
```

```
return $rez/($n-1);
}
function m2(&$v,&$u){
 $rez=0;
 $n=count($v);
 for($i=1;$i<$n;$i++)
  $rez+=$v[$i]*$u[$i];
 return $rez/($n-1);
}
```

Spearman correlation coefficient

The statistics of Spearman rank correlation coefficients are computed as follows:

- Spearman correlation coefficient:

    $r_sp = coef_rk($poz[0],$poz[1]);

where *$poz[0]* is the position on sort series of measured inhibitory activity, and *$poz[1]* is the position on sort serried of estimated inhibitory activity by MDF-SAR model with four descriptor.

- t Student parameter:

    $t_sp = t_p($n,1,pow($r_sp,0.5));

- Significance of t Student parameter

    $p_sp = p_t($t_sp,$n-2);

Semi-quantitative correlation coefficient

The statistics of semi-quantitative correlation coefficients are computed as follows:

- Semi-quantitative correlation coefficient:

    $r_sq = pow($r_pe$r_sp,0.5);*

- t Student parameter:

    $t_sq = t_p($n,1,pow($r_sq,0.5));

- Significance of t Student parameter

    $p_sq = p_t($t_sq,$n-2);

Results

$$r_{Spr}^2 = 0.8606$$

$$t_{Spm,1} = 20.03 \tag{2}$$

$$p_{Spm,1} = 1.62 \cdot 10^{-29}$$

$$r_{sQ}^2 = 0.8829$$

$$t_{sQ} = 22.14 \tag{3}$$

$$p_{sQ} = 5.57 \cdot 10^{-32}$$

Kendall's rank correlation coefficients

Definition

Kendall-tau is a non-parametric correlation coefficient that can be used to assess and test correlations between non-interval scaled ordinal variables. Frequently the Greek letter τ (tau), is use to abbreviate the Kendall tau correlation coefficient.

The Kendall tau correlation coefficient is considered to be equivalent to the Spearman rank correlation coefficient. While Spearman rank correlation coefficient is like the Pearson correlation coefficient but computed from ranks, the Kendall tau correlation rather represents a probability.

There are three Kendall's tau correlation coefficient known as tau-a, tau-b, and tau-c.

Formula

Let $(Y_{m-i},\ Y_{est-i})$ and $(Y_{m-j},\ Y_{est-j})$ be the pair of measured and estimated inhibitory activity. If $Y_{m-j} - Y_{m-i}$ and $Y_{est-j} - Y_{est-i}$, where $i < j$ have the same sign the pair is *concordant*, if have opposite signs the pair is *discordant*.

In a sample of *n* observations it can be found *n(n-1)/2* pairs corresponding to choices $1 \leq i < j \leq n$.

The formulas of Kendall's tau correlation coefficients are as follows:

- Kendall tau-a correlation coefficient ($\tau_{Ken,a}$):

$$\tau_{Ken,a} = (C-D)/[n(n-1)/2]$$

- Kendall tau-b correlation coefficient ($\tau_{Ken,b}$):

$$\tau_{Ken,b} = (C-D)/\sqrt{[(n(n-1)/2-t)(n(n-1)/2-u)]}$$

where t is the number of tied Y_m values and u is the number of tied Y_{est} values.

- Kendall tau-c correlation coefficient ($\tau_{Ken,c}$):

$$\tau_{Ken,c} = 2(C-D)/n^2$$

Interpretation

- If the agreement between the two rankings is perfect and the two rankings are the same, the coefficient has value 1.
- If the disagreement between the two rankings is perfect and one ranking is the reverse of the other, the coefficient has value -1.
- For all other arrangements the value lies between -1 and 1, and increasing values imply increasing agreement between the rankings.
- If the rankings are independent, the coefficient has value 0.

Related statistics

- $\tau_{Ken,a}^2$ = the squared of Kendall tau-a correlation coefficient;
- $\tau_{Ken,b}^2$ = the squared of Kendall tau-b correlation coefficient;
- $\tau_{Ken,c}^2$ = the squared of Kendall tau-c correlation coefficient.

Statistical significance

Statistical significance of the Kendall's tau correlation coefficient is testes by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Kendalls tau correlation coefficients was:

- Kendall tau-a correlation coefficient:

  H_0: $\tau_{Ken,a} = 0$ (there is no correlation between the two variables)

  H_1: $\tau_{Ken,a} <> 0$ (the two variables are correlated)

- Kendall tau-b correlation coefficient:

  H_0: $\tau_{Ken,b} = 0$ (there is no correlation between the two variables)

  H_1: $\tau_{Ken,b} <> 0$ (the two variables are correlated)

- Kendall tau-b correlation coefficient:

  H_0: $\tau_{Ken,c} = 0$ (there is no correlation between the two variables)

  H_1: $\tau_{Ken,c} <> 0$ (the two variables are correlated)

PHP implementation

Kendall function was implemented in order to calculate the Kendall's tau correlation coefficients:

function Kendall(&$cmp){

```
$n = count($cmp[0]);
$pz = 0;
if(!is_numeric($cmp[0][0])) $pz = 1;
$C = 0;
$D = 0;
$E = 0;
for($i=$pz;$i<$n-1;$i++)
        for($j=$i+1;$j<$n;$j++){
                $sgx = 0;
                $sgy = 0;
                if($cmp[0][$i]>$cmp[0][$j]) $sgx = 1;
                if($cmp[0][$i]<$cmp[0][$j]) $sgx = -1;
                if($cmp[1][$i]>$cmp[1][$j]) $sgy = 1;
                if($cmp[1][$i]<$cmp[1][$j]) $sgy = -1;
                if($sgx*$sgy>0) $C++;
                if($sgx*$sgy<0) $D++;
                if($sgx*$sgy==0) $E++;
                if($sgx==0)$tied_x[$i][]=$j;
                if($sgy==0)$tied_y[$i][]=$j;
        }
$t1 = 0;        $u1 = 0;
$vt = 0;        $vu = 0;
$v2t = 0;       $v2u = 0;
if(isset($tied_x))
if(is_array($tied_x)){
        foreach($tied_x as $vx){
                $nt = count($vx)+1;
                $t1 += $nt*($nt-1);
                $vt += $nt*($nt-1)*(2*$nt+5);
                $v2t += $nt*($nt-1)*($nt-2);
        }
}
if(isset($tied_y))
if(is_array($tied_y)){
        foreach($tied_y as $vy){
                $nu = count($vy)+1;
                $u1 += $nu*($nu-1);
                $vu += $nu*($nu-1)*(2*$nu+5);
                $v2u += $nu*($nu-1)*($nu-2);
        }
}
$v1 = $t1*$u1;
$t1 /= 2;
$u1 /= 2;
$v2 = $v2t*$v2u;
$S = $C - $D;
$n = $n - $pz;
$cn2 = $n*($n-1)/2;
```

```
$tau_a2 = pow($S,2)/pow($cn2,2);
$v_tau_a = $cn2*(2*$n+5)/9;
$z_tau_a = $S/pow($v_tau_a,0.5);
$T = ($cn2-$t1)*($cn2-$u1);
$tau_b2 = pow($S,2)/$T;
$vT0 = $v_tau_a - ($vt + $vu)/18;
$vT1 = $v1/(4*$cn2);
$vT2 = $v2/(18*$cn2*($n-2));
$v_tau_b = pow($vT0 + $vT1 + $vT2 , 0.5);
$z_tau_b = $S/$v_tau_b;
$gamma = pow(($C - $D)/($C + $D),2);
$v_gamma = (2*$n+5)/9.0/$cn2;
$z_gamma = $gamma/pow($v_gamma,0.5);
$tau_c2 = 4*pow($S,2)/pow($n,4);
$z_tau_c = $z_tau_b*($n-1)/$n;
    return array( $tau_a2, $z_tau_a, $tau_b2, $z_tau_b,
    $tau_c2, $z_tau_c, $gamma, $z_gamma );
    }
```

where C is the number of concordant pairs (C = (<, <) or (>, >)), D is the number of discordant pairs (D = (<, >) or (>, <)), and E is the number of equal pairs (E = (=, .) or (., =)).

Results

- Kendall's τ_a correlation coefficient and associated statistics:

$$\tau_{Ken,a}{}^2 = 0.6129$$
$$Z_{Ken,\tau a} = 9.37 \qquad\qquad (4)$$
$$p_{Ken,\tau a} = 7.44 \cdot 10^{-21}$$

- Kendall's τ_b correlation coefficient and associated statistics:

$$\tau_{Ken,b}{}^2 = 0.6177$$
$$Z_{Ken,\tau b} = 9.37 \qquad\qquad (5)$$
$$p_{Ken,\tau b} = 7.26 \cdot 10^{-21}$$

- Kendall's τ_c correlation coefficient and associated statistics:

$$\tau_{Ken,c}{}^2 = 0.5948$$
$$Z_{Ken,\tau c} = 9.23 \qquad\qquad (6)$$
$$p_{Ken,\tau c} = 2.70 \cdot 10^{-20}$$

Gamma correlation coefficient

Definition

The Gamma correlation coefficient (Γ, gamma) is a measure of association between variables that comparing with Kendall's tau correlation coefficients is more resistant

to tied data [25], being preferable to Spearman rank or Kendall tau when data contain many tied observations [26].

Formula

The formula for Gamma correlation coefficient is:

$$\Gamma = (C-D)/(C+D)$$

where the significance of C and D were described above.

Interpretation

- In the same manner as the Kendall tau correlation coefficient.

Related statistics

- Γ^2 = the squared of Gamma correlation coefficient.

Statistical significance

Statistical significance of Gamma correlation coefficient was tested by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Gamma correlation coefficients was:

H_0: $\Gamma = 0$ (there is no correlation between the two variables)

H_1: $\Gamma <> 0$ (the two variables are correlated).

PHP implementation

The function which computes the Gamma correlation coefficient was presented at Kendall's tau correlation coefficient, in PHP implementation section.

Results

$\Gamma^2 = 0.6208$

$Z_\Gamma = 7.43$ (7)

$p_\Gamma = 1.11 \cdot 10^{-13}$

[25] Goodman L. A., Kruskal W.H., *Measures of association for cross-classifications III: Approximate sampling theory*, J. Amer. Statistical Assoc. 58, 1963, p. 310-364.

[26] Siegel S., Castellan N. J., *Nonparametric Statistics for the Behavioural Sciences*, 2nd Edition, McGraw-Hill, 1988.

Conclusions

All seven computational methods used to evaluate the correlation between measured and estimated by MDF-SAR model inhibitory activity are statistically significant (p-value always less than 0.0001, correlation coefficients always greater than 0.5).

More research on other classes of biologic active compounds may reveal whether it is appropriate to analyze the MDF-SAR models using the Pearson correlation coefficient or other correlation coefficients (Spearman rank, Kendall's tau, or Gamma correlation coefficient).

Auto-calibrated Online Evaluation: Database Design and Implementation

Abstract

An auto-calibrated online evaluation system for student knowledge assessment was developed and has been implemented at the Faculty of Materials Science and Engineering, the Technical University of Cluj-Napoca. The system is available for students' end-of-course evaluation on physical chemistry, microbiology and toxicology, and materials chemistry topics and it can be used at the test center. The auto-calibrated online evaluation methodology, databases and program implementation, computer-system design and evaluation are reported.

Keywords

Auto-calibrated online evaluation, Computer-system design, Multiple choice questions (MCQs), Evaluation methodology, Databases and program implementation

Introduction

Educational measurement at the university level has been moved in the last years from the paper-and-pencil testing towards the use of computer- and/or Internet-based testing. Computer-based testing refers to performing examinations via stand alone or network computers [27]. The examination questions are displayed on a computer screen and the student choose their answers using the computer's mouse or keyboard; the responses are recorded by the computer.

---

[27] Olsen J. B. [online], *Guidelines for computer-based testing*. Retrieved May/June 2000. Available at http://www.isoc.org/oti/printversions/0500olsen.html.

Computer-based tests can be found at all educational levels (pre-university [28] and university [29]) and in many universities all over the world (Lagos state University College of Medicine, Nigeria [30]; Heriot-Watt University, Edinburgh, Scotland [31]; University of Hertfordshire, United Kingdom [32, 33]; Loughborough University, Leicestershire, U. K. [34]; Pennsylvania State University, U. S. A. [35]; University of Toronto, Ontario, Canada [36];

---

[28] Ananda A. L., Gunashingham H., Hoe K. Y., Toh Y. F., D*esign of an intelligent on-line examination system*, Computers & Education, 13(1), 1989, p. 45-52.

[29] Frosini G., Lazzerini B., Marcelloni F., *Performing automatic exams*, Computers & Education, 31(3), 1998, p. 281-300.

[30] Oyebola D. D., Adewoye O. E., Iyaniwura J. O., Alada A. R., Fasanmade A. A., Raji Y., *A comparative study of students' performance in preclinical physiology assessed by multiple choice and short essay questions*, African journal of medicine and medical sciences, 29(3-4), 2000, p. 201-205.

[31] Beevers C. E., McGuire G. R., Stirling G., Wild D. G., *Mathematical ability assessed by computer*, Computers & Education, 25(3), 1995, p. 123-132.

[32] Lilley M., Barker T., Britton C., *The development and evaluation of a software prototype for computer-adaptive testing*, Computers & Education, 43(1-2), 2004, p. 109-123.

[33] Barker L. M., Britton C. T., *Automated Feedback for a Computer-Adaptive Test: A Case Study*, Proceedings for 9th CAA Conference 2005 [online] [cited 13 May 2006], Available   at http://www.caaconference.com/pastConferences/2005/proceedings/LilleyM_BarkerT_BrittonC.pdf

[34] Croft A. C., Danson M., Dawson B. R., Ward J. P., *Experiences of using computer assisted assessment in engineering mathematics*, Computers & Education, 37(1), 2001, p. 53-66.

[35] Ariew R., *A diagnostic test for students entering a computer-assisted learning curriculum in French*, Computers & Education, 3(4), 1979, p. 331-333.

[36] Brydges R., Classen R, Larmer J., Xeroulis G., Dubrowski A., *Computer-assisted assessment of one-handed knot tying skills performed within various contexts: a construct validity study*, The American Journal of Surgery, 192(1), 2006, p. 109-113.

University of Southern Queensland, Queensland, Australia [37]; University of Hawai'i at Manoa [38]; Kulak University, Belgium [39]). Many researchers compared the equivalence of computer-based and paper-and-pen tests and most of them conclude that computer may be use in many traditional multiple-choice test settings without any significance on student performance [40, 41, 42]. Some differences were observed on tests which include extensive reading passages, when the students proved to obtained lower performance on computer-based tests comparing with on paper tests [43, 44].

---

[37] Taylor A. J., *Self Test: a flexible self assessment package for distance and other learners*, Computers & Education, 31(3), 1998, p. 319-328.

[38] Brown J. D., *Computers in language testing: present research and some future directions*, Language Learning & Technology, 1(1), 1997, p. 44-59.

[39] Demeersseman E., Wylin B., Panen J., *TOETS: development of a computer assisted assessment system*, Proceedings for 5th CAA Conference 2001 [online] [cited 13 May 2006] Available at
http://www.caaconference.com/pastConferences/2001/proceedings/n1.pdf

[40] Peak P., *Recent Trends in Comparability Studies* [online]; Released August 2005 [cited July 2006] [thirty-four pages], Available at
http://www.pearsonedmeasurement.com/downloads/research/RR_05_05.pdf

[41] Russell M., Goldberg A, O'Connor K., *Computer-Based testing and Validity: A Look Bach and Into the Future* [online], Released August 2003 [cited July 2006], Available at:
http://www.bc.edu/research/intasc/PDF/ComputerBasedValidity.pdf

[42] MacCann R., *The equivalence of online and traditional testing for different subpopulations and item types*, British Journal of Educational Technology, 37(1), 2006, p. 79-91.

[43] Mazzeo J., Harvey A. L., *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (ETS RR-88-21)*, Princeton, NJ: Educational Testing Service, 1988.

[44] O'Malley K. J., Kirkpatrick R., Sherwood W., Burdick H. J., Hsieh M.C., Sanford E.E., *Comparability of a Paper Based and Computer Based Reading Test in Early Elementary Grades*, Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada, April, 2005.

Starting from the experiences obtained by creation of the multiple-choice examination system for general chemistry topic [45, 46], the present paper concerned with issues involved in creation of an auto-calibrated online system, computer-system designs, database and program implementation, and system evaluation.

Material and Method

Multiple choice questions methodology

An important part of a computer-based testing is represented by creation of the multiple-choice questions (MCQs), which contain the procedure of creation, storing and managing of MCQs.

The anatomy of multiple-choice test contains two basic parts: a statement or a situation, a problem (*question* or *steam*) and a list of suggested solutions (*alternatives* or *options*). The steam may be construct in form of a question or of an incomplete statement and the list of options must always contains at least one correct or best alternative one and a number of incorrect options (*distractors*). The distracters must appear as *plausible* solutions to the problem for the students who have not achieved the interest objective measured by the test and *implausible* solutions for those students who have achieved the objective (only the correct option(s) should appear plausible to these students).

Regarding the anatomy of the questions, each problem has five suggested options and must test comprehension, application as well as students' analysis abilities. Regarding the number of correct options, the multiple choice questions cover the followings classes:

- **Single correct answer**: all except one of the options are incorrect, the remaining option is the correct answer;

- **Best-answer**: the alternatives differ in their degree of correctness. Some options may be completely incorrect and some incompletely incorrect, but at least one option is completely correct;

[45] Naşcu H. I., Jäntschi L., *Multiple Choice Examination System 1. Database Design and Implementation for General Chemistry*, Leonardo Journal of Sciences, 5, 2004, p. 18-33.

[46] Naşcu H. I., *Jäntschi L., Multiple Choice Examination System 2. Online Quizzes for General Chemistry*, Leonardo Electronic Journal of Practices and Technologies, 5, 2004, p. 26-36.

- *Multiple responses*: two or more of the options but never all five options are keyed as correct answer. These king of questions can be scored in different ways, but most frequently use are:
  - o *All-or-none rule*: one point if all the correct answers and none of the distractors are selected, and zero points otherwise. This was the alternative choused for our system;
  - o *Scoring each alternative independently*: one point for each correct answer chosen and one point for each distractor not chosen.

Methodology of scores and final mark

In order to provide an objective evaluation the scores and final mark methodologies take into consideration multiple factors some of them directly related with students knowledge and some related with associated activities.

There were implemented two types of scores which are directly related with students acquired knowledge: the score of an individual test and average scores of student tests. The score of an individual test characteristics and modality of computing are presented in table 1.

In order to compute the final mark, since each student had the possibility of testing his/her knowledge as many time as he/she want in the imposed period of examination, individual average score is computed. In computing of the individual average score, all student scores were taken into consideration with one exception. If a student had at least two scores, the lowest one is deleted and the left scores are used to compute his/her average testing score.

Table 1. Scores and mark methodologies

Parameter	Description
tmn_rc_global	The average time needed to give a correct answer (take into consideration all tests stored into database)
tm_rc	The average time needed to give a correct answer for the current test (just the data from the current test are taken into consideration)
C_t	The time coefficient Formula: $C_t = tm\_rc / tmn\_rc\_global$
$C_{c\text{-}a}$	The coefficient of correct-answered questions Formula: $C_{c\text{-}a} = nr\_rc/nr\_rc\_global$
m_g	The geometric mean Formula: $m_g = sqrt(C_t * C_{c\text{-}a})$
$T_{score} = m_g * 10$	The test score Formula: $T_{score} = m_g * 10$

The final mark can took values from 4 (the lowest average of test scores; the exam is failed) to 10 (the higher average of test scores). The final mark methodology took into

consideration the students scores as well as his/her accomplish (bonus, up to 2-3 points) or non-accomplished (penalty, subtracted up to 1 point) of assumed activities (e.g. stems banking construction activities). There are applied penalties (subtracted 0.5 point) every time when the student gives up after beginning of a test.

Examination methodology

The computer-based examination methodology imposes the followings:

- Place of examination: at test center, C414 room, the Technical University of Cluj-Napoca;

- Type of examination: computer- and teacher-assisted. This modality was choused in order to avoid cheating and copying. The teacher was responsible for login of each student to the computer-based testing interface;

- Period and time of examination: imposed by the academic year structure (period of examination; one month) and by the teacher (time of examination; were defined for each day from the week the hours when the lab was open for computer-based test examination);

- Number of MCQs: thirty;

- MCQs test generation: each test was automated generated by the application, using a double randomization method. The first randomization was applied in chousing the stems for each test and the second one for displaying the options order. The second randomization assured the randomization of correct options (it is known that there is a tendency to make "B", "C", and/or "D" the correct choices and to avoid "A" and/or "D" [47]);

- Chousing the correct answers: by mouse selection;

- Number of computer-based tests: as many as the students want;

- Test results: auto-calibrated of scores and final mark according with data stored into database every time when a student performed a new test.

Database design

The implementation of computer-based testing system was performed by the use of a relational database `*chemistry*` that store multiple choice banking (`…_tests`), students

---

[47] Atkinson M., *Writing Effective Multiple Choice Exams*, The Successful Professor 1(6), 2002, p. 2-4.

information (`..._users`) and information regarding the evaluations (`mcqs_results`). The MySQL database server [48] stores the database on vl.academicdirect.ro server, AcademicDirect domain.

The structure of the table that store the multiple choice banking (`..._tests`) is in table 2. In table 2 are specified the name of the field, its type and description of the field.

Table 2. The structure of the `..._tests` table

Field name	Field type	Specification
`id`	bigint(20)	An auto increment value which is manipulated by database server and serves as primary key for record unique identification
`name`	varchar(30)	Store the name of the user which create the steam
`item`	varchar(255)	The steam
`o1`	varchar(255)	First possible option
`o2`	varchar(255)	Second possible option
`o3`	varchar(255)	Third possible option
`o4`	varchar(255)	Forth possible option
`o5`	varchar(255)	Fifth possible option
`a1`	tinyint(4)	Correctness (`1`) or incorrectness (`0`) of the first possible option
`a2`	tinyint(4)	Correctness (`1`) or incorrectness (`0`) of the second possible option
`a3`	tinyint(4)	Correctness (`1`) or incorrectness (`0`) of the third possible option
`a4`	tinyint(4)	Correctness (`1`) or incorrectness (`0`) of the forth possible option
`a5`	tinyint(4)	Correctness (`1`) or incorrectness (`0`) of the fifth possible option

The structure of the table that store the students information (`..._users`) is in table 3.

Table 3. The structure of the `..._users` table

Field name	Field type	Specification
`id`	tinyint(4)	Auto increment value that serves as primary key for record identification
`name`	varchar(20)	Store user second and first name
`p`	varchar(255)	Store the penalties apply to the student (if were apply)
`bonus`	float	Store the bonus points (if were apply)
`pass`	varchar(32)	Store encrypted password

The structure of the table that store the students information (`mcqs_results`) is in table 4.

The `..._tests` table are related with `..._users` and respectively `mcqe` tables.

Table 4. The structure of the `mcqs_results` table

Field name	Field type	Specification

[48] MySQL [online], MySQL AB, ©1995-2006 [cited 2006 June], Available at http://www.mysql.com

`id`	bigint(20)	Auto increment value that serves as primary key for record identification
`subj`	varchar(10)	Store the topic abbreviation (there were included into the system MQCs for three topics: physical chemistry (*phys_chem*), microbiology and toxicology (*microb_tox*), and materials chemistry (*mate_chem*)
`name`	varchar(30)	Store the names of the student that performed the test
`suid`	varchar(32)	Store encrypted password
`qlist`	varchar(255)	Store the list of the questions that appeared on the test
`tb`	varchar(17)	Store the dates (dd.mm.yy hh.mm.ss format) corresponding with the start of examination
`te`	varchar(17)	Store the dates (dd.mm.yy hh.mm.ss format) corresponding with the end of examination
`p`	tinyint(4)	Store the number of obtained points (the number of correct answers)

System interface

PHP (Hypertext Preprocessor) [49] was use to implement the auto-calibrated online evaluation system. The characteristics of the main programs are in Table 5.

Table 5. Auto-calibrated online evaluation environment: programs characteristics

Program name	Function	Remarks
index.php	Main window of auto-calibrated online evaluation system	The main interface contains a description of the computer-bases test environment, of scores and respectively final mark methodologies
wrong_questions.php	Display the wrong questions stored into database	There are display the name of the author, the stem, the options and the correctness or incorrectness of each option
index_add.php	Allows submission of a new MCQ and/or correction of a MCQ stored into database	The interface allows: • to go to *new_user.php* interface; • to add new stems and options and to define the options correctness or incorrectness; • displaying of previous saved MCQs. The changing of the stems and associated options is password protected (just the user which create the MCQ are allowed to change it); • to link towards special symbols and/or characters.
new_user.php	Allows registration of a new user	
tests.php	Testing interface	The access to this interface is password protected (just the teacher as administrator had the password)
test_points.php	Display scores and final mark	
definitions.php	Securities specifications	

[49] PHP [online], The PHP Group, ©2001-2006 [cited 2006 June], Available at http://www.php.net.

Results and Discussions

The auto-calibrated online evaluation system was created and it is available via the address: http://vl.academicdirect.org/general_chemistry. Until now, the system was use in order to test students' knowledge at the Faculty of Materials Science and Engineering, the Technical University of Cluj-Napoca, Romania regarding following topics:

- Physical Chemistry: third year of study;
- Microbiology and Toxicology: forth year of study;
- Materials Chemistry: first year of study.

The access to the system is restricted by checking the IP addresses, being available just at the test center.

MCQs banking

Three MCQs banking were created, one for each above presented topics. The total numbers of MCQs stored into databases according with each topic are: 424 for Physical Chemistry topic, 363 for Microbiology and Toxicology topic and 865 for Materials Chemistry topic.

The distribution of the stems with one, two, three and for correct options according with each topic, express as relative frequency (f_r (%)) and associated 95% confidence interval (95% CI) compute by the use of online software which use the Newcomb method with continuity correction [50] are in Table 6.

Table 6. MCQs banking characteristics according with topics

Correct options	Physical Chemistry		Microbiology and Toxicology		Materials Chemistry	
	f_r (%)	95%CI	f_r (%)	95%CI	f_r (%)	95%CI
One	49.3	[44.4, 54.2]	65.3	[60.1, 70.1]	34.4	[31.3, 37.7]
Two	26.9	[22.8, 31.4]	16.3	[12.7, 20.5]	23.2	[20.5, 26.2]
Three	16.5	[13.2, 20.5]	10.5	[7.6, 14.2]	20.9	[18.3, 23.8]
Four	7.31	[5.1, 10.3]	7.99	[5.5, 11.4]	21.4	[18.7, 24.3]
Total	100		100		100	

Analyzing the results from Table 6 it can be observed that the number of MCQs with one correct option is almost fifty percent on physical chemistry topic and exceed the half from the total number of questions for microbiology and toxicology topic (65.3%). The

---

[50] Newcombe R. G., *Two-Sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods*, Statistics in Medicine, 17, 1998, p. 857-872.

heterogeneous distribution of the questions with one, two, three and respectively four correct options could be a disadvantage of the MCQs banking on physical chemistry and microbiology and toxicology topics. Looking at the MCQs banking on materials chemistry topic it can be observed that there is a homogeneous distribution of questions with one, two, three, and respectively four correct answers. Thus, the proportion of MCQs with one correct answer is a little bit greater comparing with the proportions of MCQs with two, three and four correct answers, but the differences are significant.

What can be done in the future regarding the MCQs banking? On physical chemistry, respectively microbiology and toxicology topics, through creation of another questions, the distribution of the items with two, three and four correct options can become homogeneous.

Computer-assisted tests environment

The main interfaces of auto-calibrated online evaluation environment for all three topics are similar and look like in figure 1.

Figure 1. The main interface of the auto-calibrated online evaluation system

The system allows management of MCQs on predefined topics. In order to introduce a MCQ into database, the user must to register and to choose a password. When the user decide to introduce MCQs into database, by choosing `Add/Modify Questions` link the system

allows to chouse the user name by selecting it from a drop-down list and open the interface of adding new questions and/or modifying previous saved questions (see figure 2). In modifying interface, first all MCQs created by the user are displayed and by selecting the stem `id` which the user want to change it, the user can change the selected stem, associated options and definition of options correctness or incorrectness. A modification of a steam and/or associated options is password protected, just the user which created and saved the question having the right to change it.

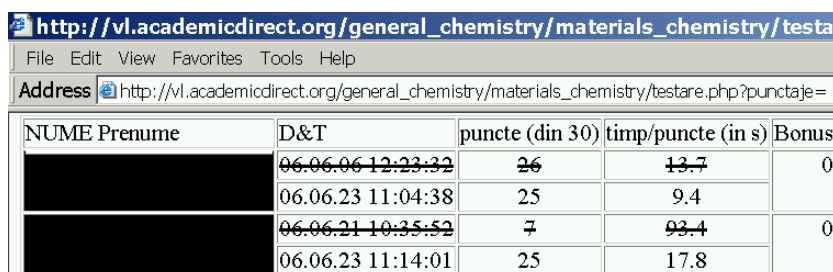*Figure 2. Interface for modifying of a MCQ*

The testing is imposed to take place at the test center and the access to it is protected by teacher password. The teacher is the one which open the test for each student. For each student, the system display automated on the screen thirty MCQs, each student having a different set of questions. Each steam has five possible options (the number of correct options is not specified). On the left side of each option, there is a radio-button that allows selection of the correct answers.

The finalization of the test by selecting of the `I finished the test` button did not allow returning at the testing interface and display on the screen the identification data of the student (first and second name), the time when the test was started and when was ended and the number of correct answers from a maximum of thirty.

From the main interface of the system there is access to the obtained points and marks of all students which performed the exam on one of the three specified topics (`Obtained Points` link, figure 1). The results are display as data organized in tables and graphical representations (distribution of the obtained scores and distribution of the final marks).

As results, the data are organized into three tables:

- First table: contains student identification details, date and time of the examination, the number of obtained points, the time necessary to give a correct answer, and the bonus points when was applicable (see figure 2);
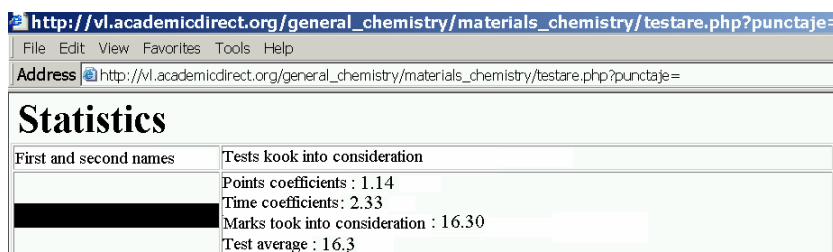
Figure 3. General results of computer-base testing

- Second table: contains details regarding marks and averages of testing with specification of the average time needed to give a correct answer and the average of the correct answers. (see figure 3);

Figure 4. Detailed statistics of tests results

- Third table: contains final statistics with specification of the correspondence between the points and the mark equal with 4 and between the points and the mark equal with 10 and the final results (student identification data, final score, the bonus points, and the final mark) (see figure 3).

Figure 5. Final results

The advantages offered by the developed online evaluation system are:

- Questions and their options are carefully structured. In these conditions the students could not guess the correct answers by applying a process of elimination;

- Questions have one or up to four correct options and the number of correct answer(s) is not specified;

- Each student has a different individual test. In this condition, the problem of copying among students is eliminated. Randomization of items and of associated options prevent learning of the correct options associated to an item (for example: if the student learned that for item no. 412 the correct options are "A", "C", and "D" it is sure that on his/her test the order of the options will not be the same) and promote learning of knowledge;

- The analysis of each student is performed through an auto-calibrated approach each time when a student performed the test, being automatically, in real time. Thus, at the end of examination each student can see his/her performances.

The system has also its disadvantages. Being an online evaluation system required access to computers and to a local-network. The system also required as prerequisite from the students' minimal knowledge of using a computer. In order to enable students to familiarize with the new evaluation environment, before the final examination, pre-exam evaluations

were open to be performed at the examination center by any interested students and as many times as they wanted.

Conclusion

The auto-calibrated online evaluation system is a valid and reliable solution in students' knowledge evaluation, providing a multi criteria objective assessment environment.

Communication of Results on Risk Factors Studies: Confidence Intervals

Abstract

Purpose: Starting from the hypothesis that confidence intervals are used in medical research as a criterion of trustworthiness and robustness of findings, the aim of the research was to determine whether the medical parameters communicated as results in abstracts of risk factors studies published in PubMed database and in Romania online journals are accompanied by confidence intervals.

Method: The search strategy included four keywords and some limitations (publication data, the language of publication, and studies on human). Four inclusion criterions were imposed. The obtained results were summarized and analyzed with Statistica.

Results: A number of 3191 were identified after applying the search strategy on PubMed database. Almost three and a half percent (n = 110) remained after inclusion the confidence intervals as keyword. Sixty articles out of one hundred and one accomplished the inclusion criteria. A number of sixty-five articles from nine Romanian online journals were identified. After applying the inclusion criterions, five articles out of sixty-five were included into the analysis. Testing the null hypothesis that there are not significant differences between the number of articles indexed in PubMed database and the number of articles published in Romanian online journals which refer as outcome the relative risk or odds ratio with associated confidence intervals, a p-value less than 0.0001 was obtained.

Conclusion: It can be conclude that the publication standards in Romanian medical journals must by arise in order to become aligned with the international trends and standards.

Keywords

Risk Factors, Cohort Study, Relative Risk, Odds Ratio, Confidence Intervals (CIs)

Introduction

The inclusion of the best available evidence into clinical decision is the main goal of evidence-based practice [51]. In the spirit of evidence-based practice, the physicians must to be able to translate the knowledge resulted from medical research by including them into

[51] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. Br Med J 1996;312(7023):71-72.

individual decisions [52,53]. In the assessment of an article, the physicians must deal with fundamental statistics such as the standard deviation, the standard error and the confidence intervals (CIs) [54].

Confidence intervals defines as *an estimated range of values that is likely to include an unknown population parameter, the estimated range being calculates from a given set of sample data* [55] are used nowadays as a criterion of assessment of the trustworthiness or robustness of the finding [56]. If independent sample are take repeatedly from same population, and the confidence interval is calculated for each sample, then a certain percentage (called confidence level) of the interval will include the unknown population parameter. Confidence intervals are usually computed for the 95%. The CIs offer to the physicians the possibilities to be more certain about the clinical value of a statistical parameter and to decide on what degree he/she can rely on the results [54].

Starting from the hypothesis that confidence intervals and their widths are used in medical research as a criterion for results trustworthiness and robustness, the objective of the research was to determine whether the medical parameters communicated in abstracts as results of risk factors studies published in PubMed database and in Romania online journals are accompanied by associated CIs.

Material and Method

Screening for risk factors studies were search using PubMed electronic database for articles in English and Romanian online journals for articles in Romanian.

---

[52] Kitson A, Harvey G, McCormack B. Enabling the implementation of evidence based practice: a conceptual framework. Qual Health Care. 1998;7(3):149-58.

[53] Sussman S, Valente TW, Rohrbach LA, Skara S, Pentz MA. Translation in the health professions: Converting science into action. Eval Health Prof 2006;29(1):7-32.

[54] Medina SL, Zurakowschi D. Measurement Variability and Confidence Intervals in Medicine: Why Should Radiologist Care? Radiology 2003;226:297-301.

[55] Huw D. What are confidence intervals? What is…. Hayward Group Publication. 2003;3:1-9.

[56] Bender R, Berg G, Zeeb H. Tutorial: Using confidence curves in medical research. Biom J 2005;47(2):237-47

The following keywords were used in searching strategy: screening AND relative risk OR odds ration OR cohort OR/AND confidence intervals. In searching strategy the following limits were imposed:

- Dates: 2003, February - 2006, February;
- Human or animals: Humans;
- Languages: English;
- Type of article: Clinical Trial, Editorial, Meta-Analysis, Review, Guideline

For inclusion into the study the abstract of the reported research had to be a screening and had to have:

- The abstract available;
- Data concerning risk factors;
- The results express as relative risk or odds ration;
- A cohort design of the research.

The confidence intervals keyword it was choused to be optional (OR/AND) because in the Romanian journals the confidence intervals it is not as frequently used as into the PubMed indexed journal. Also, the type of article was not imposed in searching the Romanian online journals.

Studies identified through the above search strategy that met the inclusion criteria were included into the analysis. There were collected the following data: (1) the name of statistical parameter, (2) the confidence intervals associated with the parameter (as dichotomial variable: Yes/No), (3) the abbreviation of the journal, (4) publication type.

Data were summarized and analyzed with Statistica 6.0 software. The 95% CIs for proportions were calculated based on the binomial distribution hypothesis [57].

Results

There were identified a number of 3191 after applying the search strategy on PubMed database. Almost three and a half percent (n = 110) remained after inclusion of confidence intervals keyword. Sixty articles out of one hundred and one accomplished the inclusion

---

[57] BinomialDistribution. ©2005, Virtual Library of Free Software, AcademicDirect [cited March 2006]. Available from: URL:

http://vl.academicdirect.org/applied_statistics/binomial_distribution/

criteria. Five articles refer as results both statistical parameters (relative risk and odds ratio). The distributions of the type of statistical parameter and associated confidence intervals, and the type of article according with the journal abbreviation are presented in Table 1.

The first step of searching the Romanian journal databases was represented by identification of online journals. There were identified a number of ninety-five Romanian medical journals. Sixty-three of them (66.31%, 95%CIs [55.80-75.78]) did not had an online version, eleven (11.22%, 95%CIs [6.33-19.99]) were accessible online after payment of a fees, and five (5.26%, 95%CIs [2.12-11.57]) had un-functional addresses.

The second step was represented by application of the search strategy. There were identified sixty-five articles from nine journals (*Stetoscop*, *Cercetări experimentale şi medico-chirurgicale*, *Timişoara Medical Journal*, *Romanian Journal of Internal Medicine*, *Observatorul Medical*, *Infomedica*, *Clujul Medical*, *Brain Aging International Journal*, *Acta Medica Transilvanica*, and *Revista Română de Psihiatrie*). Thirty-five percent did not refer any statistical parameter. Fifty-eight percent (95% CIs [46.18-70.75]) refer as results the following statistical parameters: the ratio (thirteen articles), the risk of progression (one article), the individual risk (one article), the prevalence (five articles), the p-value (one article), the mortality (one article), the mean (one article), the frequency (three articles), and the likelihood ration (one article). After applying of the inclusion criterions, five articles out of sixty-five were included into the analysis (95% CIs [3.10- 16.90]) (see

Table 2).

Table 1. Type of statistical parameter according with type of article: PubMed database

Journal abbreviation	Statistical parameter			Type of article
	RR	OR	CIs	
Acad Emerg Med	1		1	Controlled Clinical Trial
Am Heart J	1		1	Multicenter Study
Am J Cardiol	1		1	Multicenter Study
Am J Obstet Gynecol		1	1	Clinical Trial
Am J Prev Med		1	1	Clinical Trial
Arch Gen Psychiatry		1	1	Meta-Analysis
Arch Neurol	1		1	Multicenter Study
Atherosclerosis	1		1	Clinical Trial
Birth Defects Res A Clin Mol Teratol	1		1	Multicenter Study
Breast	1	1	n.a.	Review
Can J Surg		1	1	Meta-Analysis
Cancer Causes Control		1	1	Multicenter Study
Cancer Epidemiol Biomarkers Prev	1	1	2	Clinical Trial
Cancer Sci	1		1	Clinical Trial
Cardiol Young	1		1	Multicenter Study
Chest		1	1	Meta-Analysis
CMAJ	1	1	1	Clinical Trial
Cochrane Database Syst Rev		12	10	5 Meta-Analysis + 5Review
Crit Care Med	1	1	2	Clinical Trial, Multicenter Study
Drugs Aging	1		1	Clinical Trial
Eur J Cancer		1	1	Clinical Trial
Exp Mol Med	1		1	Multicenter Study
Fertil Steril	1		1	Multicenter Study
Hernia		1	1	Multicenter Study
Int J Cancer	1	1	2	Clinical Trial + Multicenter Study
J Am Soc Nephrol		2	1	Clinical Trial + Review
J Card Surg		1	1	Meta-Analysis
J Clin Oncol		2	1	Multicenter Study + Meta Analysis
J Epidemiol Community Health	1		1	Clinical Trial
J Fam Pract		1	1	Clinical Trial
J Hypertens		1	1	Multicenter Study
J Natl Cancer Inst	2	2	4	2 Meta-Analysis + Multicenter Study + Randomized Controlled Trial
J Viral Hepat		1	n.a.	Review
J Womens Health (Larchmt)	1	1	1	Multicenter Study
Neurology		1	1	Guideline
Obstet Gynecol		3	3	2 Clinical Trial + Meta-Analysis
Pediatr Infect Dis J	1		1	Multicenter Study
Saudi Med J		1	1	Multicenter Study
Scand Cardiovasc J	1		1	Multicenter Study
Sex Transm Dis	1		1	Multicenter Study
Thromb Haemost		1	1	Meta-Analysis
Total	23	42	56	1 Controlled Clinical Trial; 8 Review; 19 Multicenter Study; 1 Guideline 15 Clinical Trial; 13 Meta-Analysis

RR = Relative Risk; OR = Odds Ratio; CIs = Confidence Intervals; n.a. = not available

Table 2. Type of statistical parameter according with type of article: Romanian online journals

Journal title	Statistical parameter			Type of article
	RR	OR	CIs	
Stetoscop	2		1	Medical News + Clinical Guideline

Experimental & Medical-Surgical Researches	2	1	n.a	2 Original Article
Total	4	1	1	1 Medical News 1 Clinical Guideline 2 Original Articles
RR = Relative Risk; OR = Odds Ratio; CIs = Confidence Intervals; n.a. = not available				

Fifty-six out of sixty included articles from PubMed database refer as outcome relative risk or odds ration and associated confidence intervals (95%CIs [83.36-98.31]). One out of five articles from Romanian online journals refers as outcome relative risk with associated confidence intervals (95%CIs [4.0-76.0]).

Testing the null hypothesis that there are not significant differences between the number of articles indexed in PubMed database and the number of articles published in Romanian online journals which refer as outcome the relative risk or odds ratio with associated confidence intervals, a p-value less than 0.0001 was obtained.

Discussion

In specialty literature there is a trend which impose specification of the confidence intervals associated with statistical parameters, being considered as useful tools in study results interpretations and in generalizing of results [58]. Thus, confidence intervals are found in study of assessment of a new therapeutic attitude [59,60], the effectiveness and usefulness of a diagnostic test [61,62], in epidemiological studies [63], as well as in research which present the

---

[58] Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986;292(6522):746-750.

[59] Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: confidence intervals. CMAJ 1995;152(2):169-173.

[60] Scheen AJ. Assessment of therapeutic efficacy, an essential step in evidence-based medicine. Rev Med Liege 2000;55(4):206-210.

[61] Newcombe RG. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Stat Med 1998;17:857-872.

[62] Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol 1991;44:763-770.

[63] Bewick V, Cheek L, Ball J. Statistics review 11: assessing risk. Crit Care 2004;8(4):287-91.

evaluation of risk factors [64,65]. The main question at which the confidence intervals respond in medical studies is "What is the range of real effects that are compatible with used data?". When 95% confidence intervals are used, these confidence intervals will contain in 95% of the time the true value of the outcome effect (eg. relative risk, odds ratio, etc). In interpretation of confidence intervals for relative risk or odds ratio, the findings are non-significant if the confidence intervals embraces the value of no effect, which in these cases is equal with one. The most important reasons of using confidence intervals associated with relative risk and odds ration in risk factors studies are as follows: (1) emphasized precision of estimation; (2) its meaning is reasonable clear for any physicians (there is no confusion between significance and magnitude as in p-value interpretation); (3) nowadays journals require them.

Analyzing the results of study two remarks can be made. First remark refers the use of confidence intervals associated with relative risk or odds ration in risk factor assessments. Ninety-three percent from articles indexed in PubMed and published between 2003, February and 2006, February proved to take into consideration the necessity of confidence intervals. Opposite, twenty percent from the articles published in Romanian online journals provide the confidence intervals associated with relative risk, even if the necessity of using the confidence intervals was first published in specialty literature in 1986 [58]. In reality, the differences are bigger, because the sample size of articles published in PubMed is twenty-two time bigger comparing with the sample size of articles published in Romanian online journals. Second remark is indirect related with the objective of the study and refers the type of articles in which the statistical parameters of interest were used. Looking at the results of articles indexed in PubMed, it can be observed that the majority of the articles were multicenter studies (eighteen articles), followed by the clinical trials (fifteen articles) and meta-analysis (thirteen articles) (Table 1). Regarding the type of articles, there could not be made any comparisons because the types of articles published in Romanian online journals are not the

---

[64] Zhu K, Hunter S, Payne-Wilks K, Sutcliffe C, Bentley C, Roland CL, Williams SM. Potential differences in breast cancer risk factors based on CYP1A1 MspI and African-American-specific genotypes. Ethn Dis 2006;16(1):207-215.

[65] Bosetti C, Gallus S, La Vecchia C. Aspirin and cancer risk: an updated quantitative review to 2005. Cancer Causes Control 2006;17(7):871-888.

same as that indexed in PubMed, and on the other hand the number of articles is to less (two original articles, one clinical guideline, and one medical news, see

Table 2).

There are some explanations regarding the results of the study. The absence of the confidence intervals associated with relative risk or odds ratio in the articles published into Romanian online journals could be due to unfamiliarity of Romanian researcher with confidence intervals. The best solution for dismiss the unfamiliarity with confidence intervals could be short training on basic statistics for physicians interested on the subject or a didactic articles on confidence intervals. Another dimension of the problem refers the absence of the instruments which to compute the confidence intervals for relative risk and odds ratio. Even if for example SPSS and Statistica software did not provide these types of calculations, the physicians could use EpiInfo free software, which is able to compute based on a normality distribution the confidence intervals for relative risk and odds ratio.

Based on this study, it can be conclude that the publication standards in Romanian medical journals must by arise in order to become aligned with the international trends and standards. The request of confidence intervals associated with statistical parameters had to be imposed by the medical journals editors and Romanian authors must to comply and to include them in original articles.

Computer-Based Testing on Physical Chemistry Topic: A Case Study

ABSTRACT

According with national trends in the objective evaluation of undergraduate students' knowledge, an auto-calibrated online evaluation system was developed. The aim of the research was to assess the knowledge on physical chemistry topic of the undergraduate first year students' at the Faculty of Materials Science and Engineering, the Technical University of Cluj-Napoca, Romania by the use of the developed auto-calibrated system. The methodology of multiple-choice questions construction and the evaluation methodology are presented. The students performances in terms of number of correct answers and time needed to give a correct answer were collected and analyzed. The future plans of system development are highlighted.

KEYWORDS

auto-calibrated online evaluation, multiple choice questions (MCQs), physical chemistry, undergraduate students

INTRODUCTION

In universities, the cardinal premise of the end-of-course examination is to assess as objective as possible the students' knowledge and skills acquired on the courses, practical activities and seminars.

Development of communication (Valcke & De Wever 2004) and information technologies (Kidwell, Freeman, Smith & Zarcone 2004; Matusov, Hayes & Pluta 2005) provide today the opportunity of creation interactive computer-assisted environments used in many domains, including in chemistry training (Beasley 1999; Frecer, Burello & Miertus 2005; Chen, Chen & Cao 2002) and evaluation (Timmers, Baeyens, Remon & Nelis 2003; Stewart, Kirk, LaBrecque, Amar & Bruce 2006).

In many academic domains, educational measurement has been moving towards the use of computer-based testing, define as tests or assessments that are administered by computer in

either stand-alone or dedicated network, or by other technology devices linked to the Internet or the World Wide Web most of them using multiple choice questions (MCQs). The computer-assisted evaluation strategies are used in medicine (Oyebola, Adewoye, Iyaniwura, Alada, Fasanmade & Raji 2000), chemistry (Ananda, Gunasingham, Hoe & Toh 1989), language testing (Brown 1997), biology (Evans, Gibbons, Shah & Griffin 2004), computer science (Barker & Britton 2005), and economics (Judge 1999).

Currently in Romania, at the Faculty of Materials Science and Engineering, the Technical University of Cluj-Napoca, the traditional method (a combination of essay examination, practical examination and/or tutor assessment) is the most frequently used as evaluation of students' knowledge. In the last years, the number of students increased and the conventional examination method become time consuming in term of the examination time as well as in term of papers assessment. Thus, the students received their final marks beyond the day of examination (the next day after examination in the best case and up to one week or more). A solution of examination in large classes of students is an automated testing system which to allow testing the students knowledge and displaying immediate on the screen the examination results.

According with national and international trends in objective undergraduate students' knowledge evaluation, and starting from the experiences obtained by creation of the multiple choice examination system for general chemistry topic (Naşcu & Jäntschi 2004a; Naşcu & Jäntschi 2004b), an auto-calibrated online evaluation environment was developed (Jäntschi & Bolboacă 2006). The aim of the research was to study students' knowledge on physical chemistry topic by the use of the developed auto-calibrated online evaluation system.

MATERIAL AND METHOD

Auto-calibrated online evaluation system

The auto-calibrated online evaluation system (Jäntschi & Bolboacă 2006) embodies:

- Multiple-choice banking. The characteristics of the MCQs and respectively of the multiple-choice banking are as follows:
  - The question anatomy: a statement or a situation, a problem (*steam*) and a list of five suggested solutions (*options*). Each question had one or up to four correct options;

- o The students enrol voluntarily in the team responsible by creation of the items banking (a team of two students was responsible for creation of MCQs from the material presented at one course or one practical activity);
- o A number of four hundred and twenty-four MCQs were included into database: 49.3% with *one correct option*, 26.9% with *two correct options*, 16.5% with *three correct options*, and 7.31% with *four correct options*;
- o Score methodology: *all-or-none rule* (one point if the correct answer is selected (for questions with one correct option) or if all the correct options (respectively two for the questions with two, three for the questions with three, and four for the question with four correct options) and none of the distracter(s) - the incorrect option(s) presented as a choice in a multiple-choice test - are selected, and zero points otherwise).
- Testing environment. The online testing environment embody:
  - o A description of the testing methodology, with following specifications: the location of the examination (at the test centre); the type of the examination (computer- and teacher-assisted); the period and the time of the examination (according with the structure of academic year and with the students and teacher program);
  - o A description of the test methodology witch contains the following specifications: the number of MCQs (thirty); the generation of the MCQs tests (double randomization from MCQs banking - randomization of the steam and randomization of the options order), the number of tests (as many times as the student wanted in the imposed period of time), the penalties (applied each time when a student give up to a begun test);
  - o A description of the scores and of the final mark methodologies;
  - o The testing environment.
- Results:
  - o The results of individual tests. At the end of each test, the students' identification data, the time when the test begun and ended and the number of correct answers are displayed. There was considered that a student give a correct answer, for a question with option A and B correct, if he/she selected both options (A and B) and did not select any other option(s).
  - o The test results. A page that contains the test results for the whole class of students, express as the number of correct answers and the time needed to give a correct answer can be visualized. It was considered a *correct answer* if for example, for a question

with three correct options all the correct options were selected and none of the two distracters.

o   The auto-calibration of the final mark. The system assigned to the lower score the mark equal with 4 and to the highest score the mark equal with 10 and place each individual score between these ranges. Each time when a student gave a test, the system auto-calibrated the final marks for all students according with the distribution of individual scores of whole students'.

Assessment of students' knowledge

There were included into the study a number of forty-two students from the Faculty of Materials Science and Engineering. The familiarization of the students with the evaluation environment was possible before the examination; the students had the possibility to use the system and to evaluate themselves as many time as he/she desired, per a period of one month.

The following variables were store into database for each evaluation: the students' identification data, the data and the time when the test begin and end (according with the yy.mm.dd hh.mm.ss format), and the number of correct answer(s) (out of thirty). There were also calculated based on stored data the individual time needed to give a correct answer (express in seconds) and the average time needed to give a correct answer (this parameter took into consideration all students).

Data were analyzed with Statistica 6.0 at a significance level of 5%. The 95% confidence intervals for proportions were calculated by the use of an original method, based on the binomial distribution hypothesis (VLFS 2005).

RESULTS

The auto-calibrated online evaluation system on physical chemistry topic was created and it is available via the address: http://vl.academicdirect.org/general_chemistry/physical_chemistry/. The access to the system is restricted (by checking the IP addresses), being available just at the imposed test location.

Each student performed at least one time the online test by the use of the auto-calibrated online evaluation system. Seventeen students out of forty-two (40.5%, 95%CI [26.25, 57.09]) were content with performances obtained at first test. The distributions of the number of tests express as relative frequency and associated 95% confidence intervals are:

- Two tests: twenty-five students' out of forty-two (59.52% [42.91, 73.75]);

- Three tests: ten students' out of forty-two (23.81% [11.96, 40.42]);

- Tour tests: five students' out of forty-two (11.90% [4.82, 26.13]);

- Five tests: two students' out of forty-two (4.76% [0.06 16.61]);

- Six and seven tests: one student out of forty-two (2.38% [0.06, 11.85]);

The intervals between first and last examination (where were applicable) varied from 1 day (minimum) to 17 days (maximum), with an average of 6 days (95%CI [3.86, 8.13]) and a median of 4 days.

Statistical characteristics associated with the number of questions at which the students gave a correct answer and the time needed to give a correct answer, according with the evaluation (1st test, … 7th test), express as average (Ave), standard deviation (StDev), mode (Mode), minimum (Min) and maximum (max) are in table 1.

Table 1. Statistical characteristics of the number of correct answers and of time needed to give a correct answer

	Test	Average [95%CI]	StDev	Mode	Median	Min	Max
Number of correct answers	1st	6.40 [5.44, 7.37]	3.09	7	6	1	17
	2nd	6.60 [5.02, 8.18]	3.83	7	7	1	14
	3rd	5.80 [2.86, 8.74]	4.10	5	5	2	15
	4th	3.60 [1.03, 6.17]	2.07	N.A.	4	1	6
	5th	5.00 [-7.71, 17.71]	1.41	N.A.	5	4	6
	6th	1.00 [N.A., N.A.]	N.A.	N.A.	1	N.A.	N.A.
	7th	2.00 [N.A., N.A.]	N.A.	N.A.	2	N.A.	N.A.
Time needed to give a correct answer	1st	214.53 [174.46, 254.60]	128.59	N.A.	173.65	50.1	682
	2nd	182.90 [121.82, 243.98]	147.96	N.A.	125.80	44	713
	3rd	93.46 [51.89, 135.03]	58.12	N.A.	71.65	31	204.8
	4th	85.98 [-11.49, 183.45]	78.50	N.A.	52.50	28.8	216
	5th	30.35 [-104.97, 165.67]	15.06	N.A.	30.35	19.7	41
	6th	65.00 [N.A., N.A.]	N.A.	N.A.	N.A.	N.A.	N.A.
	7th	52.50 [N.A., N.A.]	N.A.	N.A.	N.A.	N.A.	N.A.

N.A. = not applicable

Seventeen students were content with the performances obtained at the first test. On this sample the statistical characteristics of the number of questions at which they give correct answers and of the time needed to give a correct answer are:

- The number of correct answers: Ave = 8.76 (95% CI [7.11, 10.42]); StDev = 3.21; Mode = 7, Min = 5; Max = 17;

- The time needed to give a correct answer: Ave = 137.96 (95% CI [102.20, 173.72]); StDev = 69.55; Median = 112.4; Min = 50.1; Max = 310.4.

The performances obtained by the students which performed the test more than one time, express as number of correct answers out of thirty (n_{ca}), and time needed to give a correct answer ($t_{ca}(s)$, express in seconds) are in table 2 and 3.

Table 2. Performances of the students which performed two tests

Student id	Param	Test 1st	Test 2nd	Student id	Param	Test 1st	Test 2nd
id_01	n_{ca}	6	4	id_22	n_{ca}	4	2
	$t_{ca}(s)$	171.5	125.8		$t_{ca}(s)$	365.8	449
id_03	n_{ca}	2	7	id_24	n_{ca}	5	8
	$t_{ca}(s)$	682	152.7		$t_{ca}(s)$	195.4	112.4
id_06	n_{ca}	7	14	id_26	n_{ca}	5	12
	$t_{ca}(s)$	128.7	72.2		$t_{ca}(s)$	268	83.2
id_07	n_{ca}	4	9	id_27	n_{ca}	6	11
	$t_{ca}(s)$	387.8	119.4		$t_{ca}(s)$	134.5	70.2
id_13	n_{ca}	1	2	id_31	n_{ca}	7	14
	$t_{ca}(s)$	542	44		$t_{ca}(s)$	312.1	87
id_17	n_{ca}	3	12	id_33	n_{ca}	5	9
	$t_{ca}(s)$	321	100.8		$t_{ca}(s)$	342	100.2
id_20	n_{ca}	6	7	id_42	n_{ca}	6	7
	$t_{ca}(s)$	175.8	92.7		$t_{ca}(s)$	247.8	189.6
id_21	n_{ca}	7	8				
	$t_{ca}(s)$	150.7	86.4				

Param = parameters;
n_{ca} = number of correct answers
$t_{ca}(s)$ = time needed to give a correct answer

Table 3. Performances of the students which performed more than two tests

Student id	Param	Test 1st	2nd	3rd	4th	5th	6th	7th
id_05	n_{ca}	4	2	3	N.A.	N.A.	N.A.	N.A.
	$t_{ca}(s)$	159.30	245.00	31.00	N.A.	N.A.	N.A.	N.A.
id_09	n_{ca}	7	7	15	N.A.	N.A.	N.A.	N.A.
	$t_{ca}(s)$	170.60	205.70	40.00	N.A.	N.A.	N.A.	N.A.
id_18	n_{ca}	4	3	9	N.A.	N.A.	N.A.	N.A.
	$t_{ca}(s)$	212.80	281.00	85.30	N.A.	N.A.	N.A.	N.A.
id_35	n_{ca}	4	5	9	N.A.	N.A.	N.A.	N.A.
	$t_{ca}(s)$	414.80	396.40	127.90	N.A.	N.A.	N.A.	N.A.
id_41	n_{ca}	4	4	5	N.A.	N.A.	N.A.	N.A.
	$t_{ca}(s)$	168.00	198.80	52.20	N.A.	N.A.	N.A.	N.A.
id_16	n_{ca}	4	4	2	4	N.A.	N.A.	N.A.
	$t_{ca}(s)$	285.50	125.30	58.00	28.80	N.A.	N.A.	N.A.
id_34	n_{ca}	7	5	5	5	N.A.	N.A.	N.A.
	$t_{ca}(s)$	225.30	212.60	204.80	30.40	N.A.	N.A.	N.A.

id_39	n_{ca}	3	5	5	6	N.A.	N.A.	N.A.
	$t_{ca}(s)$	223.00	167.80	167.20	102.20	N.A.	N.A.	N.A.
id_30	n_{ca}	6	3	3	2	6	N.A.	N.A.
	$t_{ca}(s)$	244.70	141.30	110.70	52.50	19.70	N.A.	N.A.
id_19	n_{ca}	3	1	2	1	4	1	2
	$t_{ca}(s)$	136.00	713.00	57.50	216.00	41.00	65.00	52.50

Param = parameters; n_{ca} = number of correct answers;
$t_{ca}(s)$ = time needed to give a correct answer; N.A. = not applicable

In order to compare the number of correct answers gave by the students which performed the test more than one time, the Student test at a significance level of 5% was applied and the results are in table 4. The number of correct answers gave by the students at the evaluation tests was abbreviate as n_{ca-i} (where i = 1 for first test, …, i = 4 for the fourth test). There were analyzed four null hypotheses as follows:

1. There was not significant difference between the average of the number of correct answers give at the first test comparing with the second test in the sample of students which performed the test by two times (n_{ca-1st} & n_{ca-2nd});

2. There was not significant difference between the average of the number of correct answers give at the second test comparing with the third test in the sample of students which performed the test by three times (n_{ca-2nd} & n_{ca-3rd});

3. There was not significant difference between the average of the number of correct answers give at the third test comparing with the fourth test in the sample of students which performed the test by four times (n_{ca-2nd} & n_{ca-3rd});

4. There was not significant difference between the average of the number of correct answers give at the first test comparing with the last test in the sample of students which performed the test more than one time (n_{ca-1st} & $n_{ca-last}$).

Table 4. Results of comparison regarding the number of average correct answers gave by students which performed more than one test

	n_{valid}	T	p-value
n_{ca-1st} & n_{ca-2nd}	25	64.5	0.0441[*]
n_{ca-2nd} & n_{ca-3rd}	10	4	0.0910
n_{ca-3rd} & n_{ca-4th}	5	4	0.7150
n_{ca-1st} & $n_{ca-last}$	25	35	0.0017[*]

n_{ca-i} = the number of correct answers;
n_{valid} = the number of valid cases;
T = the parameter of Student test;
[*] p significant

The results of comparison regarding the time needed to give a correct answer applied to students which performed more than one test are in table 5 (where t_{ca-i} is the time needed to give a correct answer for evaluation i, where i = 1 (for 1st evaluation), …, 4 (for 4th evaluation)). There were analyzed three null hypotheses as follows:

1. There was not significant difference between the average time needed to give a correct answer at the first test comparing with the second test in the sample of students which performed the test by two times (t_{ca-1st} & t_{ca-2nd});

2. There was not significant difference between the average time needed to give a correct answer at the second test comparing with the third test in the sample of students which performed the test by three times (t_{ca-2nd} & t_{ca-3rd});

3. There was not significant difference between the average time needed to give a correct answer at the third test comparing with the fourth test in the sample of students which performed the test by four times (t_{ca-3rd} & t_{ca-4th});

Table 5. Results of comparison regarding the time needed to give a correct answer on sample of students which performed more than one test

	t_{ca}	Mean	StdDev	n_{valid}	t
t_{ca-1st} & t_{ca-2nd}	t_{ca-1st}	266.604	134.33		
	t_{ca-2nd}	182.9	147.9643	25	2.01
t_{ca-2nd} & t_{ca-3rd}	t_{ca-2nd}	268.69	174.2923		
	t_{ca-3rd}	93.46	58.11506	10	2.88*
t_{ca-3rd} & t_{ca-4th}	t_{ca-3rd}	119.64	65.67909		
	t_{ca-4th}	85.98	78.49823	5	0.62

t_{ca} = the time needed to give a correct answer; StdDev = standard deviation; n_{valid} = number of valid cases; t = parameter of Student t test; * $p < 0.05$

The average time needed to give a correct answer obtained by students at the last examination (average = 98.22 seconds, Min = 50.1 seconds, Max = 682 seconds) was significantly lower ($p = 0.000002$, $n_{valid} = 25$, see figure 1) comparing with the average time needed to give a correct answer obtained at the first evaluation (average = 266.60 seconds, Min = 19.7 seconds, Max = 449 seconds).
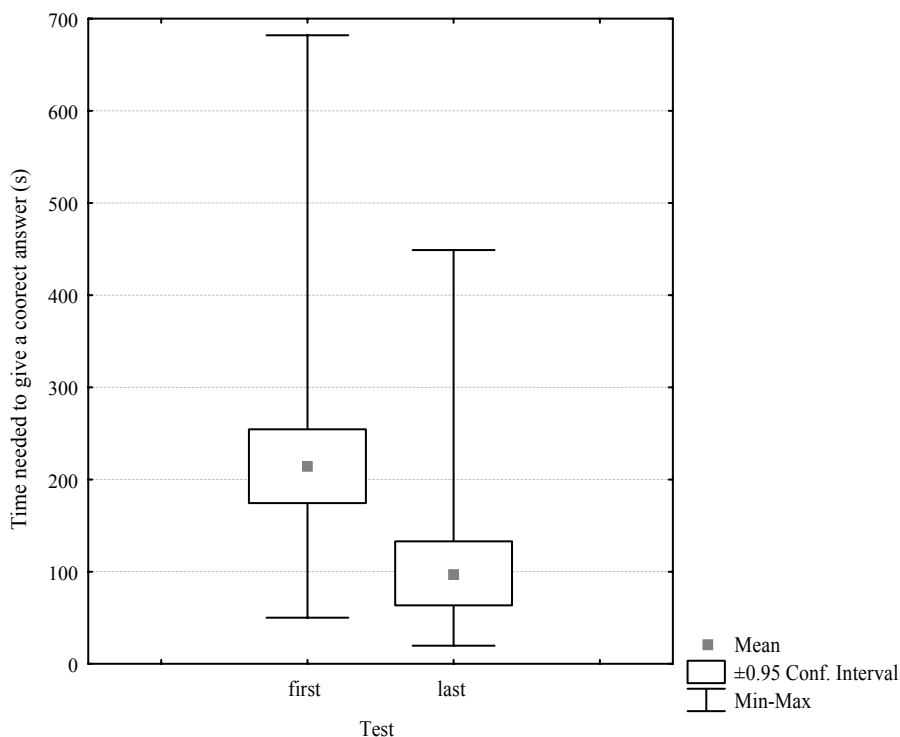
Figure 1. Distribution of time needed to give a correct answer at the first and respectively at the last test

DISCUSSIONS

The assessment of the students' knowledge is a common task at the end of the semester and/or academic year. Testing methods which imply multiple-choice questions are usually used in evaluation of students' knowledge for speed, accuracy, and fairness in grading (Toby & Plano2004).

The proposed system offer to the students involve directly in MCQs baking the opportunity to deep understand of the information regarding physical chemistry topic using an active learning method. Thus, the students were motivated to formulate questions, to create options for each question, and to define the correct answer.

Being a new evaluation method the students' had the possibility to use the system before evaluation, as pre-test evaluations. There were two aims of the pre-test evaluations. The first aim was allows familiarization of the students with proposed computer assisted evaluation environment. The second aim was to give the students possibility to test their physical chemistry knowledge, to identify their knowledge gaps, the difficult subjects and the information which need a special attention in preparation for the examination.

As it was described in Material and Method chapter, for obtaining the final mark for physical chemistry topic, each student had the possibility to test his/her knowledge as many times as desired. More than one third of students were content with performances obtained at first test. Analyzing their performances and comparing them with the whole sample, it can be observed that the average of the number of corrected answers is higher than the average reported to the whole sample, the minimum value is higher and the obtained values are not more disperse comparing with the whole sample. Comparing the average of time needed to give a correct answer, the student that decide to performed the test just one time obtained better results (137.96 seconds comparing with the whole sample, where the average was equal with 214.53 seconds). The minimum value for the variable time needed to give a correct answer was the same for the students that were content with results obtained at the first test, comparing with the rest of the sample. A significant difference was observed at the maximum value for the time needed to give a correct answer, where the value obtained by the sample of students which were content with the results obtained at the first test was half from the value obtained by the students which performed the exam more than once. This sample of students was more interested by the physical chemistry topic comparing with the colleagues who performed more than one test.

Looking at the period of time between first and last test, it can be observed the majority of students performed the test after one day, respectively three days. The students which test their knowledge after one day could by those students which learn the materials but did not try to see how the evaluation system works. Those students that performed the test again after more than one day look to be the ones which learn the materials but they were not content with obtained performances.

The results obtained at the second test, show that fifteen students out of twenty-five obtained better results in terms of number of correct answers (see table 2 and 3), and the differences vary from 1 point (id_13, id_20, id_21, id_42, id_35) to 9 points (id_17). In seventeen cases out of twenty-five, the time needed to give a correct answer decreased at the second test comparing with the first test (see table 2 and 3). The greatest decreasing was of almost 530 seconds (for the student with id_03, from 682 seconds to 152.7 seconds, table 2). These decreasing of the time needed to give the correct answer (see table 2) demonstrate that the students which presented at the second test were self-confident on their knowledge and were

able to make better connections between their acquired knowledge and the correct option(s) in less time comparing with first test.

Five out of ten students which presented to the third evaluation were able to exceed personal previous performances in terms of number of correct answers (id_09, id_18, id_35, id_39, and id_41) and of average time needed to give a correct answer (see table 3). A decrease of the average time needed to give a correct answer from 713 seconds at second test to 57.5 seconds at third test was observed at the student with id_19.

Five out of forty-two students performed the evaluation by four times. From this sample of students, three students obtained lower performances regarding the number of correct answers comparing with first evaluation, into a range from 2 points (id_34, and id_19) to 4 points (id_30). One student out of five has improved his/her performance (id_39) with two points at second evaluation comparing with first evaluation, obtaining the same performances at third evaluation as at the second evaluation, and with one point at forth evaluation comparing with third evaluation. Regarding the time needed to give a correct answer at this sample of students, except one student, the time decreased from first to forth evaluation with 121.10 seconds (id_39), 192.2 seconds (id_30), 194.9 seconds (id_34), and respectively 256.7 seconds (id_16).

Analyzing the results obtained in terms of number of correct answers and time needed to give a correct answer it can be concluded that students had improve their performances, obtaining results significantly better at the last evaluation comparing with the results obtained at the first evaluation.

Even if some students try to cheat and to obtain performances without learning the material, the auto-calibrated online system proved to be valid and did not allow or encourages these kinds of practices.

It can be conclude that the presented system is a reliable solution in students' knowledge evaluation on physical chemistry. In order to improve the auto-calibrated online evaluation system, the future direction of development has two directions: the creation of a homogenous distribution of the questions with one, two, three and respectively four correct options and the analysis of the answers gave by students to each question. The analysis of the students' answers can reveal information about the level of knowledge and will allow identification of the materials which were difficult for students to understand. With the obtain information, the practical activities, seminaries and courses on physical chemistry topic could be improving.

CONCLUSIONS

The proposed auto-calibrated online evaluation system proved to offer a stable and valid evaluation environment on physical chemistry topic.

Students' performances in terms of number of correct answers and time needed to give a correct answer reveal to be improved at final evaluation comparing with first evaluation, showing an improvement of acquired physical chemistry knowledge.

REFERENCES

1. Ananda, A. L., Gunasingham, H., Hoe, K.Y. & Toh, Y. F. (1989), "Design of an intelligent on-line examination system", *Computers and Education*, vol. 13, no. 1, pp. 45-52.

2. Barker, L. M. & Britton, C. T. (2005), "Automated Feedback for a Computer-Adaptive Test: A Case Study", *Proceedings for 9th CAA Conference 2005* [Online], viewed 12 August, 2006,
    <http://www.caaconference.com/pastConferences/2005/proceedings/LilleyM_BarkerT_BrittonC.pdf>.

3. Beasley, W. (1999), "New competencies for new times: Teacher professional development beyond 2000", *Pure and Applied Chemistry*, vol. 71, no. 5, pp. 835-844.

4. Brown, J. D. (1997), "Computers in language testing: present research and some future directions", *Language Learning & Technology*, vol. 1, no. 1, pp. 44-59.

5. Chen, C.-W., Chen, D.-Z., & Cao, G.-Z. (2002), "An improved differential evolution algorithm in training and encoding prior knowledge into feedforward networks with application in chemistry", Chemometrics *and Intelligent Laboratory Systems*, vol. 64, no. 1, pp. 27-43.

6. Evans, C., Gibbons, N. J., Shah, K. & Griffin, D. K. (2004), "Virtual learning in the biological sciences: Pitfalls of simply "putting notes on the web"", *Computers and Education*, vol. 43, no. 1-2, pp. 49-61.

7. Frecer, V., Burello, E. & Miertus, S. (2005), "Combinatorial design of nonsymmetrical cyclic urea inhibitors of aspartic protease of HIV-1", *Bioorganic and Medicinal Chemistry*, vol. 13, no. 18, pp. 5492-5501.

8. Jäntschi, L. & Bolboacă, S. D. (2006), "Auto-calibrated Online Evaluation: Database Design and Implementation", *Leonardo Electronic Journal of Practices and Technologies*, vol. 8, pp. 178-191.

9. Judge, G. (1999), "The production and use of online quizzes for Economics", *CHEER* [Online], viewed 12 August, 2006, <http://www.economicsnetwork.ac.uk/qnbank/>.

10. Kidwell, P. K., Freeman, R., Smith, C. & Zarcone, J. (2004), "Integrating online instruction with active mentoring to support professionals in applied settings", *Internet and Higher Education*, vol. 7, no. 2, pp. 141-150.

11. Matusov, E., Hayes, R. & Pluta, M. J. (2005), "Using discussion webs to Develop an academic community of learners". *Educational Technology and Society*, vol. 8, no. 2, pp. 16-39.

12. Naşcu, H. I. & Jäntschi L. (2004a), "Multiple Choice Examination System 1. Database Design and Implementation for General Chemistry", *Leonardo Journal of Sciences*, vol. 5, pp. 18-33.

13. Naşcu, H. I. & Jäntschi, L. (2004b), "Multiple Choice Examination System 2. Online Quizzes for General Chemistry", *Leonardo Electronic Journal of Practices and Technologies*, vol. 5, pp. 26-36.

14. Oyebola, D. D., Adewoye, O. E., Iyaniwura, J. O., Alada, A. R., Fasanmade, A. A. & Raji, Y. (2000), "A comparative study of students' performance in preclinical physiology assessed by multiple choice and short essay questions", *African Journal of Medicine and Medical Sciences*, vol. 29, no. 3-4, pp. 201-205.

15. Stewart, B.; Kirk, R.; LaBrecque, D.; Amar, F. G. & Bruce, M. R. M. (2006), "InterChemNet: Integrating Instrumentation, Management, and Assessment in the General Chemistry Laboratory Course", *J Chem Educ*, vol. 83, no. 3, pp. 494-500.

16. Timmers, S., Baeyens, W. R. G., Remon, J.-P. & Nelis, H. (2003), "Newer Analytical Chemistry Teaching Approaches at the Pharmaceutical Faculty of the Ghent University", *Microchimica Acta*, vol. 142, no. 3, pp. 167-175.

17. Toby, S. & Plano, R. J. (2004), "Testing, Testing: Good Teaching Is Difficult; So Is Meaningful Testing", *J Chem Educ*, vol. 81, no.2, pp. 180-181.

18. Valcke, M. & De Wever, B. (2004), "Information and communication technologies in higher education: Evidence-based practices in medical education", *Medical* Teacher, vol. 28, no. 1, pp. 40-48.

19. VLFS 2005, *Binomial Distribution*, viewed 13 August, 2006, <http://vl.academicdirect.org/applied_statistics/binomial_distribution/>.

Communication on Polynomial Formula for Vertex Cuts In B-Trees

Abstract. The paper communicates a polynomial formula giving the number and size of substructures which result after removing of one vertex from a b-tree. Particular cases of the formula are discussed.

1. INTRODUCTION

In computer science, b-trees are tree data structures that are most commonly found in databases and file systems; b-trees keep data sorted and allow amortized logarithmic time insertions and deletions [1, 2]. There are at least three domains where the b-trees concepts were use in researches:

÷ Networks: basic operations (Insert, Delete, and Search) algorithms [3, 4], dynamic collaboration [5], dynamic information storage [6], dynamic memory management [7, 8], secondary storage data structures [9], mobile databases access [10];

÷ Databases: file organization [11], access and maintain large sets of data [12,13], searching algorithms [14, 15];

÷ Computational chemistry: topological research [16], and graph theory [17,18].

It is known that connectivity is one of the basic concepts in graph theory: the minimal number of edges or vertices that disconnect a graph when removed (cuts) [19]. Why the vertex cuts are important? Vertex cuts in a graph can reveal a strong connectivity structure with better properties. The aim of the research was to found polynomial formula for vertex cuts in b-trees; was also assessed its applicability on two particular cases.

2. THE PROBLEM

A graphical representation of a b-tree is given in figure 1.
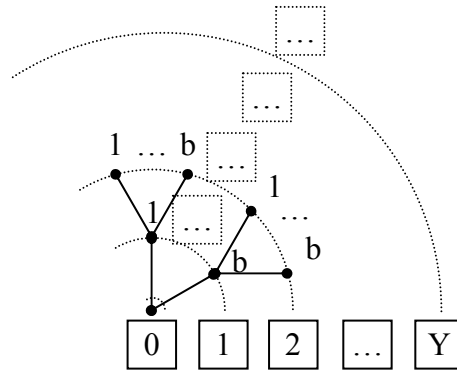
Figure 1. Tree $T_{b,Y}$

For $b = 1$ the tree degenerate into a path. For $b = 2$ the tree is the binary tree. The proposed for solving problem is counting of substructures which it results after removing of one vertex from the b-tree.

Three remarks can be making:

÷ The root vertex has b edges;

÷ The leaf vertices has 1 edge;

÷ All other vertices has $(b +1)$ edges.

3. THE SOLUTION

The total number of vertices (TNV) in a b-tree with *Y* levels where counts start from root which has assigned the level *0* (as in figure 1) is given by equation (1).

$$\left|T_{b,Y}\right| = \frac{b^{Y+1} - 1}{b - 1} \tag{1}$$

After root removing, it remains b b-trees with $|T_{b,Y-1}|$ vertices each:

$$|T_{b,Y} \setminus \text{Root}| = bX^{\frac{b^{Y}-1}{b-1}} \tag{2}$$

Leafs (one by one) removing:

$$|T_{b,Y} \setminus \text{Leaf}| = b^{Y}X^{b^{\frac{b^{Y}-1}{b-1}}} \tag{3}$$

Nodes removing (one by one, from level k, $k = 1,\ldots, Y\text{-}1$):

$$|T_{b,Y} \setminus \text{Node}_{k}| = b^{k}\left(bX^{\frac{b^{Y-k}-1}{b-1}} + X^{\frac{b^{Y+1}-b^{Y+1-k}}{b-1}} \right) \tag{4}$$

The general formula giving by the all substructures sizes and counts (ASSC) after removing one arbitrary vertex is given by equation (5):

$$\mathrm{ASSC(T_{b,Y})} = bX^{\frac{b^Y-1}{b-1}} + b^Y X^{b^{\frac{b^Y-1}{b-1}}} + \sum_{k=1}^{Y-1} b^k \left(bX^{\frac{b^{Y-k}-1}{b-1}} + X^{\frac{b^{Y+1}-b^{Y+1-k}}{b-1}} \right) \quad (5)$$

where aX^b designate a number of a connected substructures (also trees) with b vertices.

Remarks:

÷ For $Y = 0$ only the equation (1) has sense;

÷ For $Y = 1$ the equations (1), (2), and (3) should be applied;

÷ For $Y > 1$ all equations from (1) to (5) have sense and should be applied.

4. THE POLYNOMIAL FORMULA

Assigning the power of 0 at X in formula from equation 1, the polynomial formula giving the number and sizes of substructures (NSS) which it result after removing of one vertex from a b-tree can be written as:

$$\mathrm{NSS(T_{b,Y})} = \frac{b^{Y+1}-1}{b-1}X^0 + bX^{\frac{b^Y-1}{b-1}} + b^Y X^{b^{\frac{b^Y-1}{b-1}}} +$$
$$+ \sum_{k=1}^{Y-1} b^k \left(bX^{\frac{b^{Y-k}-1}{b-1}} + X^{\frac{b^{Y+1}-b^{Y+1-k}}{b-1}} \right) \quad (6)$$

Extension of node removing to $k = 0$:

$$| T_{b,Y} \setminus Node_0 | = bX^{\frac{b^Y-1}{b-1}} + X^0 = | T_{b,Y} \setminus Root | - X^0 \quad (7)$$

Extension of node removing to $k = Y$:

$$| T_{b,Y} \setminus Node_Y | = b^Y (bX^0 + X^{b^{\frac{b^Y-1}{b-1}}}) = | T_{b,Y} \setminus Leaf | - b^{Y+1}X^0 \quad (8)$$

Equations (7) and (8) allow us to rewrite the formula (6) as below:

$$\mathrm{NSS(T_{b,Y})} = \frac{b^{Y+1}-1}{b-1}X^0 - (b^{Y+1}+1)X^0 +$$
$$+ \sum_{k=0}^{Y} b^k \left(bX^{\frac{b^{Y-k}-1}{b-1}} + X^{\frac{b^{Y+1}-b^{Y+1-k}}{b-1}} \right) \quad (9)$$

Rearranging of equation (9) leads to:

$$\mathrm{NSS(T_{b,Y})} = \sum_{k=0}^{Y} b^k \left(bX^{\frac{b^{Y-k}-1}{b-1}} + X^{\frac{b^{Y+1}-b^{Y+1-k}}{b-1}} \right) - b\frac{(b^{Y+1}-2b^Y+1)}{b-1}X^0 \quad (10)$$

Remark: all equations from (6) to (10) assumes that $Y > 1$.

5. DISCUSSION OF TWO PARTICULAR CASES

The binary tree ($b = 2$) formula is obtained easily from (6) replacing b with 2:

$$NSS(T_{2,Y}) = (2^{Y+1} - 1)X^0 + 2X^{2^Y - 1} + 2^Y X^{2^{Y+1} - 2} +$$
$$+ \sum_{k=1}^{Y-1} 2^k (2X^{2^{Y-k} - 1} + X^{2^{Y+1} - 2^{Y+1-k}}) \qquad (11)$$

For $Y = 0$ (only the root is present):

$$NSS(T_{2,0}) = X^0$$

which means that no vertex cuts are available; our tree has just one vertex.

For $Y = 1$ (one root, two leafs):

$$NSS(T_{2,1}) = 3X^0 + 2X + 2X^2$$

For $Y = 2$ (one root, two nodes, four leafs):

$$NSS(T_{2,2}) = 7X^0 + 2X^3 + 4X^6 + 2(2X + X^4)$$

The unary tree (path) formula is obtained as limit formula ($b \rightarrow 1$) of (10):

$$NSS(T_{1,Y}) = \sum_{k=0}^{Y} \left(X^{Y-k} + X^k \right) - (1 - Y)X^0 \qquad (12)$$

Remark: formula (12) is according with the expected result; rearranging of (12) leads to:

$$NSS(T_{1,Y}) = 2\sum_{k=0}^{Y} X^k + (Y - 1)X^0 = 2\sum_{k=1}^{Y} X^k + (Y + 1)X^0 \qquad (13)$$

Indeed, are ($Y + 1$) vertices, and the cutting by each vertex leads to (13).

6. CONCLUDING REMARK

The obtained polynomial formulas for vertex cuts in b-trees can be generalized, allowing calculations of structures for any b and any Y. The obtained formula works also as limit formulas for trivial trees, which are paths.

ACKNOWLEDGMENT

REFERENCES

[1] B-tree definition [online]. © WIKIPEDIA - The Free Encyclopedia, 2006. [cited September 2006]. Available from: URL: http://en.wikipedia.org/wiki/B-tree (2006)

[2] R. Bayer, *Binary b-Trees for Virtual Memory*, ACM-SIGFIDET Workshop 1971, San Diego, California, **Session 5B** (1971), 219-235.

[3] D. Shasha, N. Goodman, *Concurrent Search Structure Algorithms*, ACM Transactions on Database Systems **13(1)** (1988), 53-90.

[4] H. Lu, S. Sahni, *A B-tree dynamic router-table design*, IEEE Transactions on Computers **54(7)** (2005), 813-824.

[5] B. Awerbuch, C Scheideler, *The Hyperring: A Low-Congestion Deterministic Data Structure for Distributed Environments*, Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms, **15** (2004), 311-320.

[6] E. J. A. Edemenang, E. J. D. Garba, *Dynamic information storage algorithms*. Advances in Modelling and Analysis A, **19(3-4)** (1994), 17-64.

[7] A. Laszloffy, J. Long, A. K. Patra, *Simple data management, scheduling and solution strategies for managing the irregularities in parallel adaptive hp finite element simulations*, Parallel Computing **26(13-14)** (2000), 1765-1788.

[8] J. S. Vitter, *External memory algorithms and data structures: deaimg with massive data*, ACM Computing Surveys, **33(2)** 2001, 209-271.

[9] P. Ko, S. Aluru, *Obtaining provably good performance from suffix trees in secondary storage*, Lecture Notes in Comput. Sci. **4009 LNCS** (2006), 72-83.

[10] X. Yang, A. Bouguettaya, B. Medjahed, H. Long, W. He, *Organizing and Accessing Web Services on Air*, IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, **33(6)** 2003, 742-757.

[11] Comer, *Ubiquitous b-tree*, Comput. Surv. **11(2)** 1979, 121-137.

[12] M. Schrapp, *1-Pass Top-Down Update Schemes for Search Trees. Design, Analysis and Application*, Fortschritt-Berichte der VDI-Zeitschriften, Reihe 10: Angewandte Informatik, **38** (1984), 106p.

[13] P. L. Lehman, S. B. Yao, *Efficient Locking for Concurrent Operations on b-Trees*, ACM Transactions on Database Systems, **6(4)** (1981), 650-570.

[14] T. Skopal, M. Krátký, J. Pokorný, V. Snás el, *A new range query algorithm for Universal B-trees*, Inform. Syst. **31(6)** (2006), 489-511.

[15] S.-W. Kim, *On batch-constructing B+-trees: Algorithm and its performance evaluation*, Inform. Sciences. **144(1-4)** (2002), 151-167.

[16] L.-S. Wang, S.-G. Yuan, Z. Ouyang, C.-Z. Zheng, *Important algorithms used in the target parsing system*, J. Chin. Chem. Soc. **59(2)** (2001), 241-246.

[17] M. M. Sørensen, *b-tree facets for the simple graph partitioning polytope*, J. Comb. Optim. **8(2)** (2004), 151-170.

[18] M. V. Diudea, I. Gutman, L. Jäntschi, *Molecular Topology*, 2nd Edition, Nova Science, Huntington, New York, (2002).

[19] R. Diestel, *Graph Theory*, 2nd Edition, Springer-Verlag, New York, (2000).

Subgraphs of Pair Vertices

Abstract

Subgraphs obtained by applying several fragmentation criteria are investigated. Two well known criteria (Szeged and Cluj), and two new others are defined and characterized. An example is given for the discussed procedures. The matrix and polynomial representations of vertices composing each type of subgraphs were also given. Analytical formulas for the polynomials of several classes of graphs are derived. The newly introduced subgraphs/fragments, called MaxF and CMaxF, appear to have interesting properties, which are demonstrated.

Keywords

Subgraphs; Chemical Graph Theory; Graph Polynomials

1. Definition of Fragments

Let V be a set and $E \subseteq V{\times}V$ a subset of the Cartesian product[1,2] V $\times$V. Then $G = (V, E)$ be an un-oriented graph, with $V(G)$ the set of vertices and E(G) the set of edges. A graph is connected if there is a path from one to any other vertex in G.[3,4]

Let denote by **D**(G) the distance matrix[2] of G. In terms of distance matrix, the connectivity of G is written as: $D(G)_{i,j} < \infty$ for any $i,j \in V(G)$.

The *Szeged fragmentation criterion* was introduced by Gutman.[5] Szeged subgraphs can be defined as follows:

$$SzF(G)_{i,j} = \{V(SzF(G)_{i,j}), E(SzF(G)_{i,j})\} \tag{1}$$

$$V(SzF(G)_{i,j}) = \{s \in V(G) \mid D(G)_{s,i} < D(G)_{s,j}\} \tag{2}$$

$$E(SzF(G)_{i,j}) = \{(s,t) \in E(G) \mid s,t \in V(SzF(G)_{i,j})\}. \tag{3}$$

The Szeged set $SzF(G)_{i,j}$ is a connected subgraph (the set results from a geodesic operator).

The *Cluj fragmentation criterion* was introduced by Diudea *et al.*[6-10] The Cluj subgraphs are defined on a path $p(i,j)$ separating the vertices i and j:

$$CjF(G)_{i,j,p} = SzF(G_p)_{i,j}, \tag{4}$$

where G_p is obtained from G by deleting the path p with exception of its endpoints. The Cluj fragments are also connected subgraphs [2].

A minimal subgraph of G can be defined as follows:

$$MinF(G)_{i,j} = (\{i\},\varnothing) \tag{5}$$

The above definition of minimal subgraphs is a trivial one. Such a subgraph always contains one vertex, the vertex i.

A maximal connected subgraph of G, containing the vertex i but not the vertex j, we denote here by $MaxF(G)_{i,j}$. Such maximal connected subgraphs can be constructed by using a temporary graph, $(VTemp(G)_{i,j}, ETemp(G)_{i,j})$, to be defined below, which is a disconnected graph:

$$VTemp(G)_{i,j} = \{s \in V(G) \mid s \neq j\}, ETemp(G)_{i,j} = \{(u,v) \in E(G) \mid u,v \neq j\} \tag{6}$$

The $V(MaxF(G)_{i,j})$ and $E(MaxF(G)_{i,j})$ sets are defined as follows:

$$V(MaxF(G)_{i,j}) = \{s \in VTemp(G)_{i,j} \mid D(VTemp(G)_{i,j})_{s,i} < \infty\} \tag{7}$$

$$E(MaxF(G)_{i,j}) = \{(s,t) \in E(G) \mid s,t \in V(MaxF(G)_{i,j})\} \tag{8}$$

Let now construct the complementary subgraph $CMaxF(G)_{i,j}$ of maximal connected subgraph $MaxF(G)_{i,j}$ with respect to the graph G:

$$CMaxF(G)_{i,j} = (V(CMaxF(G)_{i,j}), E(CMaxF(G)_{i,j})) \tag{9}$$

$$V(CMaxF(G)_{i,j}) = \{s \in V(G) \mid s \notin V(MaxF(G)_{i,j})\} \tag{10}$$

$$E(CMaxF(G)_{i,j}) = \{(s,t) \in E(G) \mid s,t \in V(CMaxF(G)_{i,j})\} \tag{11}$$

The new substructure $CMaxF(G)_{i,j}$ is, in general, a smaller one. Rarely it has more than one element (one vertex, the vertex j). However, it differs from $MinF(G)_{i,j} = (\{i\},\varnothing)$, as it is shown in following example:
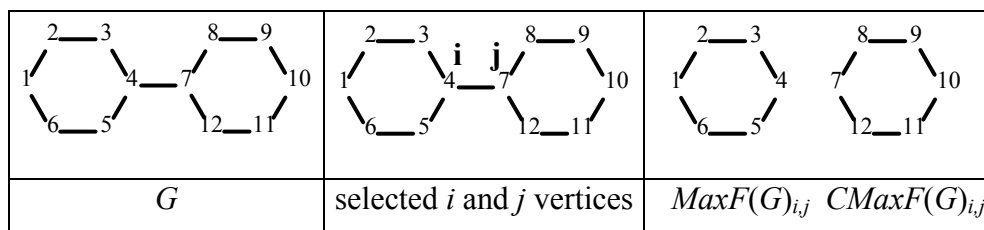

Figure 1. *MaxF(G)* and *CMaxF(G)* subgraph definition

Figure 1 illustrates a graph *G*, and the way of generating the fragments *MaxF(G)$_{i,j}$* and *CMaxF(G)$_{i,j}$*. As can be seen, *G ≠ MaxF(G)$_{i,j}$ ∪ CMaxF(G)$_{i,j}$* (by the edge (*i,j*)).

2. Matrices derived from fragmentation

Let us present the matrix representation[3,4,10] of subgraphs generated by the above four fragmentation criteria.

The entries in matrices [**MaxF**], [**CMaxF**], and [**SZ**] matrix[11] are similar and represent the vertex cardinality of subgraphs, viewed as sets of vertices:

$$[\mathbf{M}]_{i,j} = |\mathbf{M}(G)_{i,j}|, \; where \; \mathbf{M} \in \{\mathbf{MaxF, CMaxF, SZ}\} \tag{12}$$

while for **CJ** matrix (according to ref.[6,7]) is:

$$[\mathbf{CJ}]_{i,j} = \max\{|CJF(G)_{i,j,p}| \; ; \; p \in P(G)_{i,j}\} \tag{13}$$

Tables 1 to 4 give examples of the above matrices in case of graph G. Note that the above matrices are unsymmetric (exception some symmetric graphs). More about the **SZ** and **CJ** matrices the reader can find in ref.[4,10]

Table 1. Matrix **MaxF** for the graph *G*.

MaxF	1	2	3	4	5	6	7	8	9	10	11	12	
1	0	11	11	5	11	11	6	11	11	11	11	11	110
2	11	0	11	5	11	11	6	11	11	11	11	11	110
3	11	11	0	5	11	11	6	11	11	11	11	11	110
4	11	11	11	0	11	11	6	11	11	11	11	11	116
5	11	11	11	5	0	11	6	11	11	11	11	11	110
6	11	11	11	5	11	0	6	11	11	11	11	11	110
7	11	11	11	6	11	11	0	11	11	11	11	11	116
8	11	11	11	6	11	11	5	0	11	11	11	11	110
9	11	11	11	6	11	11	5	11	0	11	11	11	110
10	11	11	11	6	11	11	5	11	11	0	11	11	110
11	11	11	11	6	11	11	5	11	11	11	0	11	110
12	11	11	11	6	11	11	5	11	11	11	11	0	110
	121	121	121	61	121	121	61	121	121	121	121	121	1332

$$MaxF(G,x) = 10x^5 + 12x^6 + 110x^{11}; \; D1|_{x=1} = 1332$$

Table 2. Matrix **CMaxF** for the graph *G*.

CMaxF	1	2	3	4	5	6	7	8	9	10	11	12	
1	0	1	1	7	1	1	6	1	1	1	1	1	22
2	1	0	1	7	1	1	6	1	1	1	1	1	22
3	1	1	0	7	1	1	6	1	1	1	1	1	22
4	1	1	1	0	1	1	6	1	1	1	1	1	16

5	1	1	1	7	0	1	6	1	1	1	1	1	22
6	1	1	1	7	1	0	6	1	1	1	1	1	22
7	1	1	1	6	1	1	0	1	1	1	1	1	16
8	1	1	1	6	1	1	7	0	1	1	1	1	22
9	1	1	1	6	1	1	7	1	0	1	1	1	22
10	1	1	1	6	1	1	7	1	1	0	1	1	22
11	1	1	1	6	1	1	7	1	1	1	0	1	22
12	1	1	1	6	1	1	7	1	1	1	1	0	22
	11	11	11	71	11	11	71	11	11	11	11	11	252

$$CMaxF(G,x) = 110x + 12x^6 + 10x^7;\ D1|_{x=1} = 252$$

Table 3. Matrix **SZ** for the graph *G*.

SZ	1	2	3	4	5	6	7	8	9	10	11	12	
1	0	3	2	3	2	3	3	5	5	6	5	5	42
2	9	0	3	2	3	2	4	4	6	6	6	4	49
3	8	9	0	3	2	9	3	6	6	7	6	6	65
4	9	8	9	0	9	8	6	6	8	7	8	6	84
5	8	9	2	3	0	9	3	6	6	7	6	6	65
6	9	2	3	2	3	0	4	4	6	6	6	4	49
7	7	8	6	6	6	8	0	9	8	9	8	9	84
8	7	6	6	3	6	6	3	0	9	8	9	2	65
9	6	6	4	4	4	6	2	3	0	9	2	3	49
10	6	5	5	3	5	5	3	2	3	0	3	2	42
11	6	6	4	4	4	6	2	3	2	9	0	3	49
12	7	6	6	3	6	6	3	2	9	8	9	0	65
	82	68	50	36	50	68	36	50	68	82	68	50	708

$$SZ(G,x) = 16x^2 = 24x^3 + 12x^4 + 8x^5 + 36x^6 + 6x^7 + 12x^8 + 18x^9;\ D1|_{x=1} = 708$$

Table 4. Matrix **CJ** for the graph *G*.

CJ	1	2	3	4	5	6	7	8	9	10	11	12	
1	0	3	2	2	2	3	3	3	3	3	3	3	30
2	9	0	3	2	2	2	4	4	4	4	4	4	42
3	8	9	0	3	2	8	5	5	5	5	5	5	60
4	8	8	9	0	9	8	6	6	6	6	6	6	78
5	8	8	2	3	0	9	5	5	5	5	5	5	60
6	9	2	2	2	3	0	4	4	4	4	4	4	42
7	6	6	6	6	6	6	0	9	8	8	8	9	78
8	5	5	5	5	5	5	3	0	9	8	8	2	60
9	4	4	4	4	4	4	2	3	0	9	2	2	42
10	3	3	3	3	3	3	2	2	3	0	3	2	30
11	4	4	4	4	4	4	2	2	2	9	0	3	42
12	5	5	5	5	5	5	3	2	8	8	9	0	60
	69	57	45	39	45	57	39	45	57	69	57	45	624

$CJ(G,x) = 22x^2 + 24x^3 + 24x^4 + 24x^5 + 12x^6 + 14x^8 + 12x^9; D1\mid_{x=1} = 624$									

3. Polynomial representation of fragmentations

A counting polynomial[12-14] is a representation of a sequence of numbers, with the exponents showing the extent of partitions $p(G)$, $\cup \, p(G) = P(G)$ of a graph property $P(G)$ while the coefficients $m(G,k)$ are related to the occurrence of partitions of extent k.

$$P(G,x) = \sum_k m(G,k) \cdot x^k \qquad (14)$$

The coefficients of counting polynomials related to the above fragmentations can be obtained from the corresponding matrices, by counting of the occurrence of their entries. For the graph G, the related polynomials are given under each of the Tables 1 to 4. It is easily seen that the first derivative, in x=1, equals the sum of entries in the corresponding matrices.

The polynomial representation is useful in partitioning a graph property in view of weighting the subgraph contributions by physico-chemical characteristics of the vertices/atoms composing a molecular graph.

Table 5 lists the formulas for calculating the discussed polynomials for several classes of graphs.

Table 5. Formulas for polynomials of several classes of graphs

	G	Polynomial
1	Star $S_{1,n}$	$\text{MaxF}(G,x) = (n+1)x^0 + nx^1 + n^2 x^n$ $\text{CMaxF}(G,x) = \text{SZ}(G,x) = \text{CJ}(G,x) = (n+1)x^0 + n^2 x^1 + nx^n$
2	Path P_n	$\text{MaxF}(G,x) = nx^0 + 2\sum_{1\le k<n} kx^k$ $\text{CMaxF}(G,x) = \text{CJ}(G,x) = nx^0 + 2\sum_{1\le k<n}(n-k)x^k$ $\text{SZ}(G,x) = nx^0 + 4\sum_{1\le k\le\left(2n-1+(-1)^n\right)/4} kx^k + 2\sum_{1\le k\le\left(2n-3+(-1)^n\right)/4}(2k-1)x^{n+1-k}$
3	Complete K_n	$\text{MaxF}(G,x) = nx^0 + n(n-1)x^{n-1}$ $\text{CMaxF}(G,x) = \text{CJ}(G,x) = \text{SZ}(G,x) = nx^0 + n(n-1)x$
4	Dendrimers* $D_{f,s}$	$\text{MaxF}(G,x) = \dfrac{f(f-1)^s - 2}{f-2}x^0 + f\sum_{0\le k<s}(f-1)^{s-k-1}\dfrac{(f-1)^{k+1}-1}{f-2}x^{\frac{(f-1)^{k+1}-1}{f-2}}$ $+ f\sum_{0\le k<s}(f-1)^{s-k-1}\dfrac{f(f-1)^s-(f-1)^{k+1}-1}{f-2}x^{\frac{f(f-1)^s-(f-1)^{k+1}-1}{f-2}}$

$$CMaxF(G,x)=\frac{f(f-1)^s-2}{f-2}x^0+f\sum_{0\le k<s}(f-1)^{s-k-1}\frac{f(f-1)^s-(f-1)^{k+1}-1}{f-2}x^{\frac{(f-1)^{k+1}-1}{f-2}}$$

$$+f\sum_{0\le k<s}(f-1)^{s-k-1}\frac{(f-1)^{k+1}-1}{f-2}x^{\frac{f(f-1)^s-(f-1)^{k+1}-1}{f-2}}=CJ(G,x)$$

$$SZ(G,x)=\frac{f(f-1)^s-2}{f-2}x^0$$

$$+f\sum_{0<k\le s}\frac{(f-1)^{s-k}}{f-2}\left(\left((f-1)^{2k}-1\right)x^{\frac{(f-1)^k-1}{f-2}}+\left((f-1)^{2k-1}-1\right)x^{\frac{f(f-1)^s-(f-1)^k-1}{f-2}}\right)$$

Note: s letter in the dendrimers[15] formula is the layer; on "0" layer there is 1 atom, on "1" layer there are f atoms (counting the number of atoms $n = f + 1$).

4. Properties of subgraphs generated by fragmentation

Theorem 1. *In a not empty graph G, the following relations hold*:

$$1= n_{MinF} \le n_{CJF} , n_{SZF} \le n_{MaxF} \le n_G; \text{ where } n_{A} = |A| \tag{15}$$

Demonstration. By definition, it is immediate that $1 = n_{MinF}$; next, $n_{MinF} \le n_{CJF}$, n_{SZF} comes from: $\{i\} \in \{V(SZF(G)_{i,j}), V(CJF(G)_{i,j})\}$. The relation n_{CJF}, $n_{SZF} \le n_{MaxF}$ is true because both Szeged and Cluj fragments collect all vertices lying closer to i than to j ,excluding the vertices equidistant to i and j whereas MaxF criterion excludes, at the start, only the j vertex, and next all vertices, which become non-connected to i. These vertices are also not found in $V(CjF(G)_{i,j,p})$ and $V(SzF(G)_{i,j})$, which are both connected subgraphs containing the vertex i. The Cluj fragment excludes, in addition, the vertices belonging to the path $p_{i,j}$. The last inequality is even more immediate, as long as $V(MaxF(G)_{i,j})$ is constructed on the vertices of $V(G)$.

Even if $n_{CJF} \le n_{SZF}$, for any p from i to j, the inclusion $V(CjF(G)_{i,j,p}) \subseteq V(SzF(G)_{i,j})$ or its reverse is false, in general.

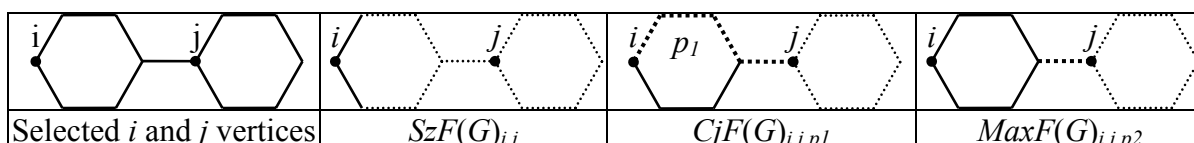The following example (Figure 2) illustrates the above results.

| Selected i and j vertices | $SzF(G)_{i,j}$ | $CjF(G)_{i,j,p1}$ | $MaxF(G)_{i,j,p2}$ |

Figure 2. *Examples of fragments MaxF(G), SzF(G)$_{i,j}$ and CjF(G)$_{i,j,p}$ subgraphs*

The *MaxF(G)$_{i,j}$* is a connected subgraph (let's denote subgraph by "$\prec$") of *G*, by construction:

$$MaxF(G)_{i,j} \prec G \tag{16}$$

Theorem 2. *CMaxF(G)$_{i,j}$ is a connected subgraph of G*:

$$CMaxF(G)_{i,j} \prec G \tag{17}$$

Demonstration. Clearly, *CMaxF(G)$_{i,j}$* is a subgraph of *G*. All its vertices and edges belong to *G*. The difficulty is to prove its connectivity. Suppose there exists a vertex *k* not connected in any way with the vertex *j* (*i.e.*, $D(CMaxF(G)_{i,j})_{k,j} = \infty$). But *k* belongs to *G*, thereby $D(G)_{k,i} < \infty$, and also $D(G)_{k,j} < \infty$. We can assume there is a path from *k* to *i*: $k = x_0, x_1, ..., x_n = i$. If none of its vertices $x_1, ..., x_{n-1}$ is *j* we goes to a contradiction because, in this case, our vertex *k* is in the wrong place (*CMaxF(G)$_{i,j}$*); its place must be *MaxF(G)$_{i,j}$*. If one of the vertices $x_1, ..., x_{n-1}$ is *j* then we have an interesting path: $k = x_0, ... x_{m-2}, x_{m-1}, x_m = j, x_{m+1} ... x_n = i$. If $(x_{m-1}, x_m) \in E(G)$, and $x_{m-1}, x_m \in V(CMaxF(G)_{i,j})$ then (by definition of *CMaxF(G)$_{i,j}$*) it results that $(x_{m-1}, x_m) \in E(CMaxF(G)_{i,j})$; thus, x_{m-1} *is connected* to $x_m = j$ in *CMaxF(G)$_{i,j}$*. By recursion to all the vertices $x_{m-2}, ..., x_0 = k$ of our path, we conclude that *k is connected to j*.

Theorem 3. *In any connected graph, with more than two vertices, the following relations hold*:

$$SzF(G)_{i,j}, CjF(G)_{i,j,p} \prec MaxF(G)_{i,j} \prec G \tag{18}$$

Demonstration comes from the definitions of the three fragments. The fact that *G* is the biggest graph is beyond dispute. If either *SzF(G)$_{i,j}$*, or *CjF(G)$_{i,j,p}$* has an edge (*s,t*), then using a similar judgment as for (17), also *MaxF(G)$_{i,j}$* must have it; also (even more simple is) for a vertex. An interesting property is revealed by the graph polynomial of *MaxF* fragmentation, by the example for the graph *G*:

$$MaxF(G,x) = 2{\cdot}5{\cdot}x^5 + 2{\cdot}6{\cdot}x^6 + 10{\cdot}11{\cdot}x^{11} \tag{19}$$

The equation (19) shows that:

Theorem 4. *The number of occurrences of a given size subgraph, generated by MaxF criterion applied to a graph G, equals the number of vertices consisting the subgraph.*

Demonstration. A subgraph *MaxF(G)$_{i,j}$*, with vertex cardinality $n_{i,j} =| MaxF(G)_{i,j} |$ refers to one and the same reference cut-point *j,* and by construction, *one fragment for each vertex i*=1, 2, …, $n_{i,j}$ can be generated, in total $n_{i,j}$ fragments.

As a corollary, a subgraph *MaxF(G)$_{i,j}$* is indexed to *each vertex i*=1, 2, …, $n_{i,j}$ or:

$$k \in MaxF(G)_{i,j} \Leftrightarrow i \in MaxF(G)_{k,j} \tag{20}$$

which is obvious. Note that *CJF* and *SZF* criteria do not induce such degeneration of fragments, due to the path exclusion and/or the equidistant vertex exclusion.

In case of *CMaxF* fragmentation, the occurrence of subgraphs comes from that of *MaxF* fragmentation. In the above example:

$$CMaxF(G,x) = 2{\cdot}5{\cdot}x^7 + 2{\cdot}6{\cdot}x^6 + 10{\cdot}11{\cdot}x^1 \tag{21}$$

It is not surprising, since, by definition, for each *MaxF(G)$_{i,j}$* a subgraph *CMaxF(G)$_{i,j}$* is generated.

Theorem 5. *For the fragments SZF, CJF, MaxF and CMaxF, the following relation is true*:

$$n_{G}{\cdot}\,n_G -1= \Sigma\, n_{MinF} \leq \Sigma\, n_{CMaxF} \leq \Sigma\, n_{CJF} \leq \Sigma\, n_{SZF} \leq \Sigma\, n_{MaxF} \leq n_{G}{\cdot}(n_G -1)^2$$

Demonstration: indeed, $n_{G}{\cdot}(n_G -1) = \Sigma\, n_{MinF}$ based on *MinF* definition; $\Sigma\, n_{MinF} \leq \Sigma\, n_{CMaxF}$ is true because always $\{i\} \in V(CMaxF(G)_{i,j})$; $\Sigma\, n_{CJF} \leq \Sigma\, n_{SZFj}$ is true because *CJF* are generated by applying *SZ* criterion on G_p, where *p* is the path, joining *i* with *j*, to be deleted from *G*; $\Sigma\, n_{SZF} \leq \Sigma\, n_{MaxF}$ because $SZ(G)_{i,j}$ is a subgraph of *MaxF(G)$_{i,j}$* (equation 18); $\Sigma\, n_{MaxF} \leq n_{G}{\cdot}(n_G -1)^2$ is true because $n_{G}{\cdot}(n_G -1)^2$ is the maximum possible value of $\Sigma\, n_{MaxF}$. It results from: the maximum size of fragment *MaxF(G)$_{i,}$*, $n_{MaxF} = n_G -1$, and next, the number of all non-diagonal entries in the matrix is $n_{G}{\cdot}(n_G -1)$, thus obtaining max $(\Sigma\, n_{MaxF}) = n_{G}{\cdot}(n_G -1)^2$.

Conclusions

Subgraphs obtained by applying several fragmentation criteria have been investigated. Two new criteria have been defined and characterized along with the well known criteria Szeged and Cluj. The matrix and polynomial representations of vertices composing each type of subgraphs have been discussed. Analytical formulas for the polynomials of several classes

of graphs have been derived. The newly introduced subgraphs/fragments, called *MaxF* and *CMaxF*, appeared to have interesting properties, which have been demonstrated.

References

1. N. Trinajstić, *Chemical Graph Theory*, IInd Ed. CRC Press, 1992.
2. F. Harary, *Graph Theory*. Addison-Wesley, Reading, MA, 1969.
3. M. V. Diudea and O. Ivanciuc, *Molecular Topology,* Comprex, Cluj, 1995 (in Romanian).
4. M.V. Diudea, I. Gutman and L. Jäntschi, *Molecular Topology*. Nova Science, Huntington, New York, 2001.
5. I. Gutman, A formula for the Wiener number of trees and its extension to graphs containing cycles, *Graph Theory Notes of New York*,1994, 27, 9-15.
6. M. V. Diudea, *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 300-305.
7. M. V. Diudea, B. Parv, and I. Gutman, *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 1101-1108.
8. M. V. Diudea, G. Katona, I. Lukovits, and N. Trinajstić, *Croat. Chem. Acta*, **1998**, *71*, 459-471
9. M. V. Diudea and O. Ursu, *Indian J. Chem.*, 42*A*, **2003**, 1283-1294.
10. M. V. Diudea, M. S. Florescu and P. Khadikar, *Molecular topology and it's applications*, EfiCon Press, 2006, Bucharest, Romania.
11. M. V. Diudea, O. M. Minailiuc, G. Katona, and I. Gutman, Szeged Matrices and Related Numbers. *Commun. Math. Comput. Chem.* (*MATCH*), **1997**, *35*, 129-143..
12. H. Hosoya, *Bull. Chem. Soc. Japan*, **1971**, *44*, 2332-2339.
13. H. Hosoya, *Discrete Appl. Math.*, **1988**, *19*, 239-257.
14. E. V. Konstantinova and M. V. Diudea, *Croat. Chem. Acta,* **2000**,*73*, 383-403.
15. M. V. Diudea and G. Katona, Molecular Topology of Dendrimers. In: Newkome, G.A. Ed., *Advan. Dendritic Macromol.* **1999**, *4*, 135-201.

Concluzii

Statistica aplicată oferă un instrument puternic de investigare pe eşantioane de populaţii. Atunci când volumul eşantionului este redus, asumpţia distribuirii acestuia după o distribuţie Gauss (normală) este eronată. Este necesar să se folosească distribuţia binomială.

Optimizarea numerică oferă un răspuns real la problema eşantioanelor distribuite binomial.

Următoarele aplicaţii oferă soluţii reale:

o calculator pentru intervalele de confidenţă la proporţii binomiale:
   http://vl.academicdirect.org/applied_statistics/binomial_distribution/

o sistem de evaluare autocalibrată online: http://vl.academicdirect.org/general_chemistry/

o sistem de educaţie medicală continuă: http://vl.academicdirect.org/medical_informatics/

o sistem de construcţie a polinoamelor caracteristice şi de numărare pe grafuri:
   http://vl.academicdirect.org/molecular_topology/graph_polynomials/

o sistem de generare a tăierilor de vârfuri în grafuri:
   http://vl.academicdirect.org/molecular_topology/vertex_cutting/

Cluj-Napoca,
la 28.10.2006

Director temă ET46/2006,
Şef. L. Dr., Ing. Lorentz JÄNTSCHI
http://lori.academicdirect.org