AFRL-SR-AR-TR-05-

0195

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for r
maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this bu .g
suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** 20 May 2005	**3. REPORT TYPE AND DATES COVERED** Final report 01 Mar 2002-28 Feb 2005

4. TITLE AND SUBTITLE Use of Biodescriptors and Chemodescriptors in Predictive Toxicology: A Mathematical/Computational Approach	**5. FUNDING NUMBERS** G AF/F49620-02-1-0138
6. AUTHOR(S) Subhash C. Basak	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Minnesota Duluth Natural Resources Research Institute 5013 Miller Trunk Hwy. Duluth, MN 55811	**8. PERFORMING ORGANIZATION REPORT NUMBER** NRRI/TR-2005/13
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of Scientific Research 4015 Wilson Boulevard Room 713 Arlington, VA 22203-1954	SPONSORING / MONITORING

NL

20050608 051

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.	**12b. DISTRIBUTION CODE**

13. ABSTRACT ***(Maximum 200 Words)***

This project focuses on a two-pronged approach to modeling toxicity data. A standard structure-based QSAR approach using chemodescriptors (descriptors based on the chemical structure of the toxicant) has been coupled with the development of biodescriptors, a novel set of mathematical descriptors derived from 2-DE proteomic gel analyses.

The research group has explored the use of chemodescriptors calculated using high-level *ab initio* quantum chemical basis sets, exploring a wide range of potential chemodescriptors as well as considering potential mechanistic approaches to chemical toxicity, e.g., using vertical electron affinity for modeling the interactions of highly reactive chemicals.

Biodescriptor development has focused on three main approaches: global descriptors that characterize the entire proteomics map, local descriptors that characterize a subset of the proteins present in the gel, and spectrum-like descriptors. These efforts have been further enhanced through the use of robust statistical approaches and the development of new statistical techniques for analyzing the full set of proteins present in a proteomics map.

14. SUBJECT TERMS Topological indices; chemodescriptors; proteomics; biodescriptors; 2-DE gel electrophoresis; QSAR; hierarchical QSAR	**15. NUMBER OF PAGES**
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	**18. SECURITY CLASSIFICATION OF THIS PAGE** UNCLASSIFIED	**19. SECURITY CLASSIFICATION OF ABSTRACT** UNCLASSIFIED	**20. LIMITATION OF ABSTRACT** UNLIMITED

I. Table of Contents

II. Objectives 3

III. Status of Effort 4

IV. Accomplishments/New Findings 4

V. Personnel Supported 13

VI. Publications 14

2002 14

2003 14

2004 15

2005 16

In press 16

Submitted 16

VII. Interactions/Transitions 17

a) Participation at Meetings 17

b) Consultative and Advisory Functions 24

c) Transitions 25

VIII. New Discoveries, Inventions, or Patent Disclosures 25

IX. Honors and Awards 25

a) Honors 25

b) Advisory/Organizational Positions Held by Dr. Basak 25

II. Objectives

The project has the following principal objectives:

a) Development of novel biodescriptors to characterize proteomics patterns (maps)

b) Use of the new biodescriptors developed in this project to predict toxicity of chemicals from their effects on proteomics patterns

c) Use of chemodescriptors alone to develop hierarchical quantitative structure-toxicity relationship (QSTR) models to predict toxicity of chemicals

d) Comparative studies of biodescriptors vis-à-vis chemodescriptors in predicting toxicity

e) Use of chemodescriptors in predicting alterations in proteomics patterns of cells as a result of exposure to chemicals

f) Development of integrated QSTR models using the combined set of chemodescriptors and biodescriptors

We will spend the major part of our project resources in the development and use of biodescriptors in predicting the toxicity of chemicals. The other important objectives will be to investigate the utility of hierarchical QSTRs using chemodescriptors alone and the combined set of biodescriptors and chemodescriptors in predictive toxicology. We plan to accomplish these objectives in terms of the following specific tasks:

1. Establishment of databases of property/activity/ toxicity of chemicals from open literature and sources from the US Air Force Research Laboratory
2. Development of novel mathematical biodescriptors from matrices associated with proteomics patterns.
3. Development of biodescriptors from embedded graphs for proteomics maps
4. Formulation of biodescriptors using direct Map to Matrix Transformation.
5. Development of information theoretic invariants from graphs of proteomics maps
6. Development of novel biodescriptors from spectrum-like representation of proteomics patterns
7. Development of biodescriptors by quantifying moments of distribution of protein spots on the proteomics maps
8. Comparison of proteomics maps using the various classes of biodescriptors
9. Utilization of biodescriptors in predicting toxicity of chemicals
10. Development and application of chemodescriptors in QSTR analysis
11. Formulation of integrated QSTRs to predict toxicity of chemicals
12. Development of novel similarity measures of chemicals from their biodescriptors and chemodescriptors
13. Analysis of the relationship of proteomics data to dosage of halocarbon applied

III. Status of Effort

At this time, Dr. Basak's research group has done extensive modeling of halocarbon toxicity using the HiQSAR approach with chemodescriptors and biodescriptors. Dr. Balasubramanian has devoted a great deal of computer time to providing high-level quantum chemical calculations included with the chemodescriptors, some of these calculations have required several days to a week per molecule. Toxicity models based on chemodescriptors are promising. Chemodescriptor-based QSARs have also been developed to predict properties related to pharmacokinetics.

Several distinct methods for characterizing proteomic maps have been developed by Drs. Basak, Hawkins, Randic, and Vracko, and have been published in peer-reviewed journals (see Publications for details). In addition, Drs. Basak, Randic, and Bajzer were invited to contribute a chapter on proteomics modeling for the book *Genomic and Proteomic Applications of Toxicity Testing*, to be published by Humana Press. The research team has also developed several pieces of software to rapidly calculate proteomics map descriptors, something that has been done by hand up to this point. Also, Dr. Hawkins has introduced two unique statistical methods for modeling halocarbon toxicity with the combined set of chemo- and biodescriptors. These approaches employ recursive partitioning and canonical clustering, and have met with some success in modeling halocarbon toxicity. Finally, work continues on the development of information-theoretic biodescriptors and moment of distribution biodescriptors. Publications on both of these topics are pending.

The research group continues its efforts to compare the validity of various proteomics map descriptors and to measure their utility vis-à-vis chemodescriptors.

IV. Accomplishments/New Findings

The major effort of this AFOSR project was focused on the development and use of biodescriptors for proteomics maps. This effort was roughly divided into two major categories: a) biodescriptors as quantifiers of complex biosignatures and b) the identification of individual protein biomarkers.

In the category of biodescriptors, three approaches have been pursued: 1) the creation of numerical invariants (or vectors of invariants) derived from embedded graphs associated with proteomics maps (2-DE gels); 2) the development of vectors derived from projections of three dimensions (protein mass, charge, and abundance) onto three planes, the (x,y), (y,z) and (x,z) planes, and 3) the development of information theoretic indices for proteomics maps based on partitioning the electrophoretic gel into nxn cells.

Results for the numerical invariants based on proteomics maps from liver tissue from rats exposed to peroxisome proliferators (PFOA, PFDA, clofibrate, and DEHP) show that the leading eigenvalue of the D/D matrix derived from embedded graphs shows reasonable power to discriminate among maps derived from mechanistically and structurally similar chemicals.

The approach derived from 3D projections led to the creation of a vector space. Euclidean distance in such a space can be used as a measure of the similarity/dissimilarity of proteomics maps. This approach has been shown to cluster peroxisome proliferators and halocarbons reasonably well.

The information theoretic approach developed recently by Basak et al partitions the entire (x,y) plane defined by protein mass and charge into a certain number of cells, nxn, where proteins in each cell are considered equivalent. Shannon's relation was then used to compute the complexity of the entire map. Preliminary results for the four peroxisome proliferators, *viz.*, PFOA, PFDA, clofibrate, and DEHP, show that this approach clusters the first three highly fluorinated and mechanistically similar chemicals together, while putting DEHP in a category by itself. Further research on the efficacy of this novel approach is in progress [Basak et al, *WSEAS* submitted; Basak et al, *JCIM* submitted].

An example of the second methodology is the association of individual protein spots with the various toxicity values of halocarbons in the primary hepatocyte. Preliminary analysis shows that a few proteins/enzymes, *e.g.,* carbonic anhydrase and the β-subunit of F1 ATPase, are strongly associated with cellular toxicity of halocarbons. Further consultation with toxicologists is in progress to determine the validity of the selected proteins as biomarkers of halocarbon toxicity.

In addition to the work on biodescriptor development, effort has been devoted to the development of chemodescriptor-based QSAR models. Chemodescriptors have also been utilized in predicting properties relevant to modeling pharmacokinetic processes, physiological-based pharmacokinetic (PBPK) modeling, and blood:brain barrier entry of chemicals. Such models will be useful in predicting the toxicity of JP-8 chemicals.

1. Biodescriptors

The development and application of biodescriptors for predictive toxicology by the NRRI research team has been expanded to incorporate three major types of biodescriptors: a) global descriptors derived from invariants of matrices associated with proteomics maps, b) a set of local invariants describing various aspects of each map (instead of one global biodescriptor), and c) spectrum-like descriptors for the characterization of proteomics patterns.

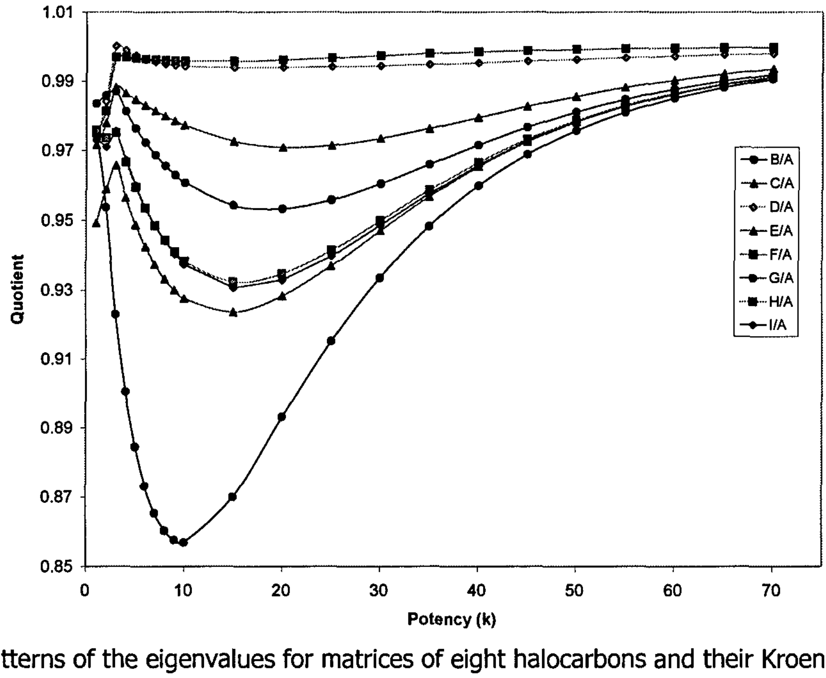

Figure 1. Patterns of the eigenvalues for matrices of eight halocarbons and their Kroenecker products.

1.1 Global Biodescriptors

Invariants of matrices, *e.g.,* the D/ D matrix, and their higher power Kronecker products have been used to develop profiles for proteomics patterns for hepatocytes exposed to eight halocarbons [Basak et al, presented at QSAR 2002]. This data was obtained from Dr. Frank Witzmann and is derived from hepatocytes exposed to halocarbons at the WPAFB, Dayton, Ohio, by Frazier, Geiss and coworkers. The profile given in Figure 1 shows that, even for the structurally-similar halocarbons, the profiles for the entire set of eigenvalues and their higher powers are non-overlapping. This indicates that these biodescriptors have no degeneracy for the set of eight compounds tested thus far. It is of interesting to see whether this pattern holds when the set of halocarbons is expanded by the addition of further data for an expanded set of halocarbons.

Basak and his collaborators at NRRI have also developed an information-theoretic parameter for the quantification of proteomics patterns. Such measures associate a number or a small set of numbers to a map to describe their relative complexity. These measures are used as biodescriptors to predict toxicity either alone or in combination with chemodescriptors. This research has only recently begun to show some results and was first presented at the 229th National ACS meeting in San Diego, March 2005. This approach divides the proteomics map even into nxn sectors, where n ≥ 2. Figure 2 shows an example of creating 2x2, 3x3, and 5x5 maps.

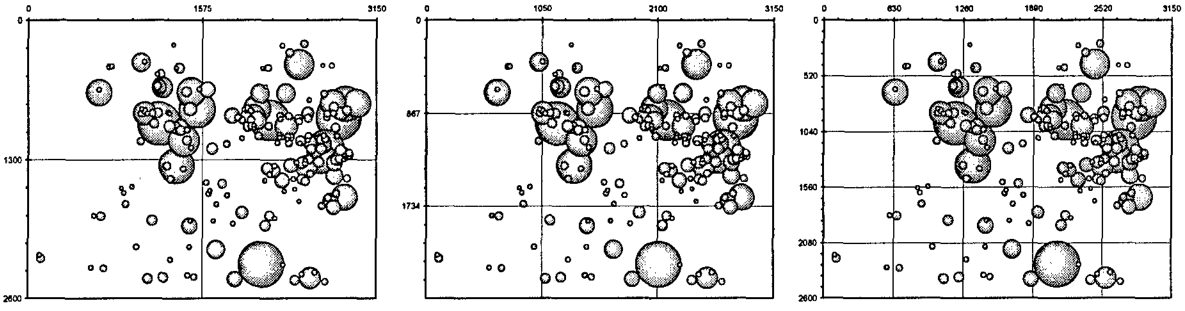

Figure 2. Division of a 200-most-abundant spot proteomics map into nxn sectors for calculation of map information content (MIC).

Once the map has been subdivided, the protein abundance in each sector was calculated and the ratio of sector abundance to total map abundance was also calculated. Using these data, a Shannon-type complexity calculation was used to find the total map complexity:

$$Complexity = -\sum_{i=1}^{h} p_i \log_2 p_i$$

These information-theoretic map descriptors have been shown to discriminate between treatments for peroxisome proliferators [Basak et al, *WSEAS* submitted; Basak et al, *JCIM* submitted]

1.2 Local Biodescriptors

Local biodescriptors are derived by examining a smaller subset of spots than those used in formulating global descriptors. Smaller subsets of *m* spots are used sequentially instead of using all *n* spots simultaneously, *m* « *n* [Randić et al, *SAR & QSAR* 2002]. The importance of local invariants in contrast to global biodescriptors is that the former indices can be used to quantify effects of selected subsets of proteins related to a particular toxicological mode of action, *e.g.*, peroxisome proliferation or apoptosis.

A two-way approach to the formulation of local biodescriptors has been developed: a) invariants derived from neighborhood graphs and b) self organizing map (SOM) analysis of protein spots to derive subsets most relevant to toxicity prediction.

In collaboration with Drs. Randic, Bajzer, Plavsic, a new kind of proteomics graph invariant has been defined which takes into consideration the local neighborhoods of spots in the maps. Such descriptors can be used in predicting toxicity and measuring similarity among proteomics maps [Randic et al, *CCA* 2004; Randic et al, *J. Proteome Res.* 2004; Randic et al, *J. Proteome Res.* submitted].

In collaboration with Dr. Vracko, neural network methods such as self organizing maps (SOM) have been used to cluster protein spots into similar groups and select a subset of spots that can be used in measuring similarity of maps and predicting toxicity of halocarbons from selected spots of maps [Vracko and Basak, *CILS* 2004]. Other measures of proteomics map similarity have been developed with Dr. Randic [Randic and Basak, *Med. Chem. Res.* in press].

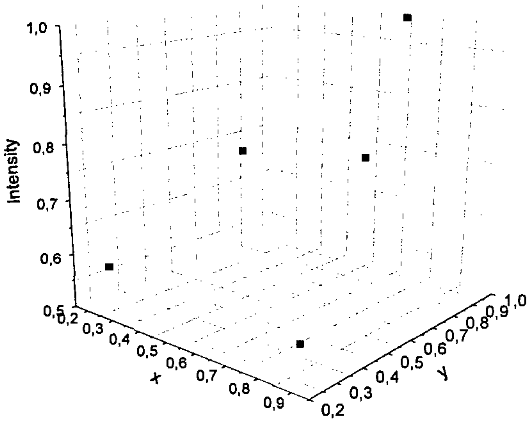

Figure 3. Five arbitrarily selected spots of reference proteomic map. The red, yellow and blue lines are projection lines on xy, xz and yz planes, respectively.

1.3 Spectrum-Like Biodescriptors of Proteomics Maps

For the purposes of developing spectrum-like biodescriptors, a proteomic map is considered to be a set of points in 3D space where each point represents a spot in the map. The x and y coordinates determine the position of the spot while the z coordinate corresponds to protein abundance (or intensity). Such a pattern of points in 3D space can be represented with three spectrum-like objects. A representation is constructed in three steps. First, the pattern is projected on the xy, xz, and yz planes, respectively (see Figure 3). In the second step, each projection (or figure) is treated separately. A figure is placed into a circle of arbitrary radius. A projection beam from the center of a circle produces a pattern of points on the circle where each point represents a particular point. In the third step, each point on the circle is taken as a center for a Lorentzian curve of the form:

$$f_i(\varphi_i) = \frac{\rho_i}{(\varphi-\varphi_i)^2+\sigma_i^2} \tag{1}$$

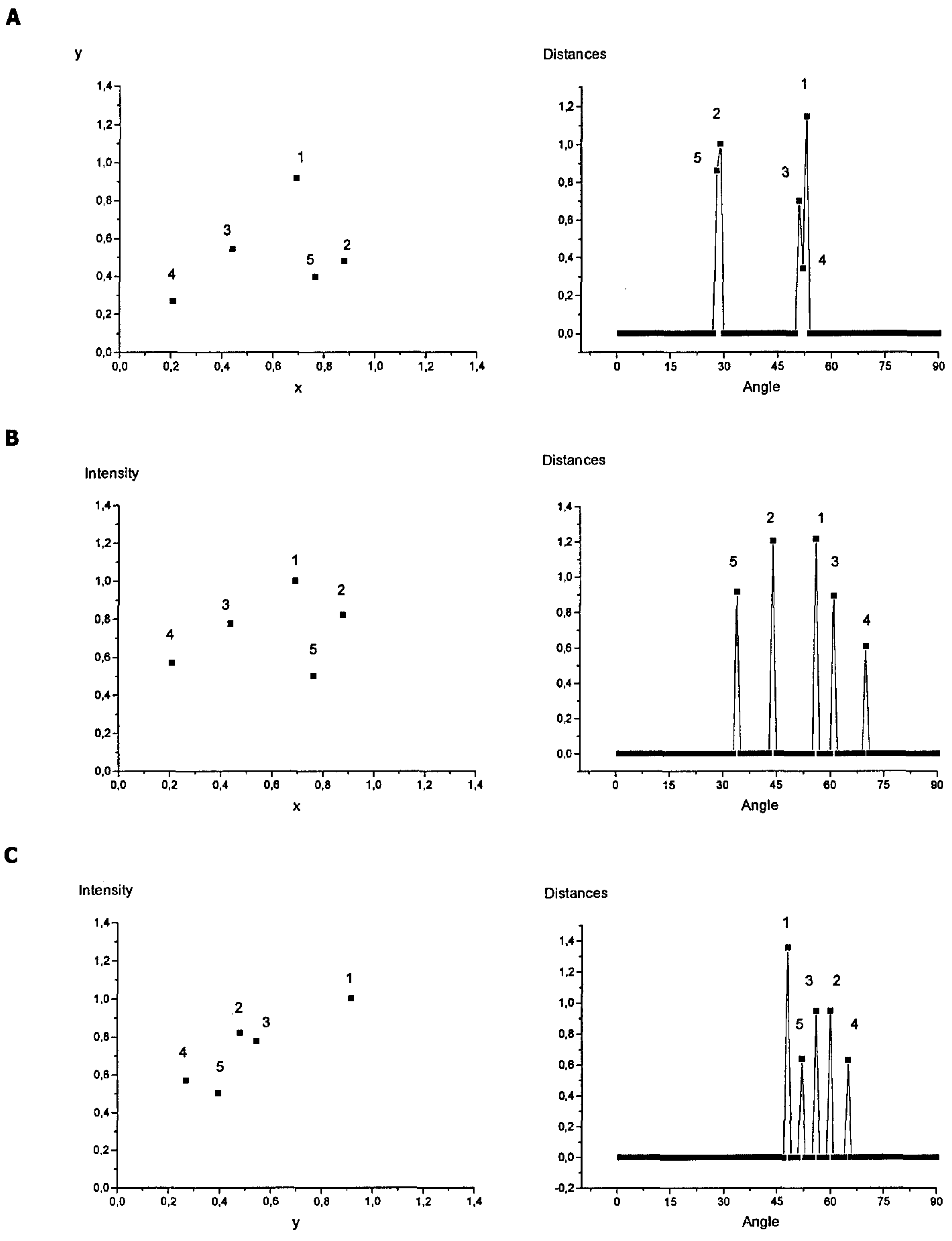

Figure 4: A: xy projection and corresponding 'spectrum-like' object
B: xz projection and corresponding 'spectrum-like' object
C: yz projection

Here, ρ_i and φ_i are the distances between the origin of the coordinate system and the position of the *i*-th point and its polar angle, respectively. σ_i is a free parameter, which can be associated with any property given to the point. If we consider only positions of proteomic peaks, the σ_i's are set to one. The spectrum related to the figure is the sum of all atomic Lorentzians and it is defined in the interval (0, 2п). An example of three projections using a pattern comprised of five points is shown in Figure 4. The complete pattern of points in the 3-D space is represented with three spectra, each of them associated with the xy, xz, or yz projection, respectively. By selecting *k* equidistant points on the interval, one figure is represented by a *k*-dimensional vector. The value of *k* determines the resolution of the representation. It is obvious that *k* should be close to the number of points in the proteomic map, otherwise some information is lost. Finally, the three spectrum-like objects are compressed into a single vector $\bar{v}$. One paper discussing this topic has already been published [Vracko and Basak, CILS 2004] and another planned for submission to the *Journal of Chemical Information and Modeling* in is preparation.

2. Chemodescriptors and Computational Toxicology

Molecular structural descriptors and topological indices, also known as chemodescriptors, have been widely used for some time now in computational toxicology and pharmaceutical design. These calculable descriptors have helped to fill in the data gap, allowing researchers to develop quantitative structure-activity relationship (QSAR) models for predicting the toxicity or activity of a candidate chemical. As part of this research effort, the computational chemistry group at NRRI has continued their efforts in developing and validating new QSAR models, as well as exploring new descriptors and novel methodologies for optimization of QSAR modeling.

2.1 Mechanism-Based HiQSARs for Predictive Toxicology

A successful HiQSAR model has been developed for a set of 55 halocarbons for which cellular level toxicity data is available [Basak et al, *JCICS* 2003]. It should be noted that this is the "superset" of compounds from which the "subset" of twenty halocarbons, tested by WPAFB and Dr. Witzmann using the DNA microarray and proteomics analysis, were selected. Selection of parameters for the HiQSAR was based on the mechanistic hypothesis that dissociative electron attachment and subsequent formation of free radicals, leading to lipid peroxidation, was a major factor in halocarbon toxicity. This conclusion was derived from previous research by Dr. Balasubramanian's, based on high-level quantum chemical calculations. If this hypothesis was correct, calculated parameters such as vertical electron affinity (VEA) should have been strongly correlated with halocarbon toxicity. Accordingly, the research team used the following classes of indices in a hierarchical manner to develop QSAR models for the toxicity of 55 halocarbons:

Topostructural (TS), topochemical (TC), geometrical (3-D), semi-empirical quantum chemical (AM1), and calculations using five *ab initio* basis sets (STO-3G, 6-31G(d), 6-311G, 6-311G* and cc-pVTZ).

Results of these analyses are summarized in Table 1 below [Basak et al, *JCICS* 2003].

It is evident from these results that only higher level *ab initio* indices (6-311G* and cc-pVTZ) made any significant improvement in model quality over topological indices for the prediction of toxicity. The lower level *ab initio* and semi-empirical indices did not make any improvement over the models developed by the topostructural and topochemical indices. It may be noted that even with the addition of these high-level *ab initio* parameters, the easily calculable topological indices outperformed all other types of molecular descriptors in predicting toxicity from structure. This shows the utility and robustness of these simple descriptors. Unfortunately, VEA was not shown to be a deciding factor in modeling halocarbon toxicity, suggesting that the mechanistic hypothesis is incorrect or that other confounding aspects of halocarbon toxicity must first be addressed.

Table 1. Summary results for the HiQSAR modeling of D_{37} in *Aspergillus nidulans* for 55 halogenated aliphatic hydrocarbons.

Model	**# indep. variables**	R^2	R^2_{cv}	*s.e.*	*F*
TSI only	2	0.3659	0.2945	1.243	15.00
TCI only	8	0.8623	0.7749	0.6161	36.00
3D only	8	0.8838	0.6496	0.5660	43.72
AM1 only	3	0.4591	0.3008	1.159	14.43
STO-3G only	3	0.3055	0.1624	1.314	7.48
6-31G(d) only	4	0.4111	0.2458	1.222	8.73
6-311G only	4	0.5140	0.3853	1.110	13.22
6-311G* only	4	0.6318	0.5053	0.9663	21.45
cc-pVTZ only	1	0.5099	0.4787	1.083	55.14
TSI + TCI	7	0.8791	0.8074	0.5710	48.83
TSI + TCI + 3D	*Same as previous model*				
TSI + TCI + AM1	7	0.8840	0.8121	0.5594	51.17
TSI + TCI + STO-3G	7	0.8724	0.7923	0.5866	45.91
TSI + TCI + 6-31G(d)	7	0.8713	0.7926	0.5893	45.45
TSI + TCI + 6-311G	10	0.9015	0.8283	0.5328	40.26
TSI + TCI + 6-311G(d)	8	0.9056	0.8335	0.5100	55.17
TSI + TCI + aug-cc-pVTZ	8	0.9476	0.9236	0.3800	103.98

2.2 Hierarchical QSAR

Dr. Basak's group has also continued its work exploring the various levels of parameters in the development of QSARs for a wide variety of properties [Basak et al, in *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens* 2003; Basak et al, *JCICS* 2003; Basak et al, *Risk Analysis* 2003; Basak et al, *Ind. J. Chem.* 2003; Basak et al, *ETAP* 2004; Gute et al, *ETAP* 2004; Hawkins et al, *ETAP* 2004; Basak and Mills, in *Chemistry of Biologically Active Synthetic and Natural Compounds* submitted; Basak et al, in *Advances in Quantum Chemistry* submitted; Basak et al, in *Biological Concepts and Techniques in Toxicology* submitted].

3. QSAR for the Estimation of Toxicokinetics-Related Properties

Collaboration with the research groups of Drs. Jim Riviere (North Carolina State University) and Jeff Fisher (University of Georgia, Athens) have presented opportunities for the NRRI research team to develop HiQSARs for problems related to JP-8 and skin toxicology, as well as toxicokinetics.

3.1 Dermal Penetration Modeling Systems

In skin toxicology, we received data from Dr. Riviere on the permeability of chemicals through membrane-coated fibers (MCFs) as models for pig skin. Preliminary results of QSAR studies on this data show that computed descriptors give reasonable models in predicting MCF data. These QSAR models will be further refined based on more extensive data being developed in Dr. Riviere's laboratory.

3.2 Development of Toxicokinetic Models

HiQSARs have been developed using calculated TS, TC descriptors in predicting blood:air, tissue:air partition coefficients [Basak, Mills et al, *SAR & QSAR* 2002; Basak et al, *Risk Analysis* 2003; Basak et al, *ETAP* 2004]. Results from these studies show that such models are comparable or superior to "mechanistic models" derived from experimental data based on an understanding of the modes of action of toxicants. It is expected that such models will be useful in providing data for PBPK modeling. Dr. Basak's research team will continue to collaborate with Dr. Jeff Fisher, Univ. of Georgia, Athens, in developing models for the prediction of toxicokinetics of JP-8 constituents as more data becomes available. Such models are expected to be very useful to the PBPK modeling community.

4. Integrated Models Using Chemo- and Biodescriptors

Basak and coworkers have formulated the idea of hierarchical quantitative structure-activity relationship (HiQSAR) models which present a minimalist approach to modeling in the sense that more expensive descriptors (demanding in computer resources or cost of laboratory testing) are only used if "less costly" predictors fail to produce adequate models. To this end, the research team has employed topological, geometrical/shape, low-level semi-empirical, and high-level *ab initio* quantum chemical descriptors to predict halocarbon toxicity. The last group of indices, calculated by Dr. Balasubramanian, can require as much as seven days of supercomputer time to calculate descriptors for one halocarbon molecule. Reasonable toxicity estimation models have been derived using such chemodescriptors [Basak et al, *JCICS* 2003; Gute et al, *ETAP* 2004]. However, in the complex domain of chemical toxicology it is possible that such calculated descriptors might not reflect important aspects of interactions relevant to toxicity. Therefore, it has been of interest to see how far supplementing the set of calculated chemodescriptors with biodescriptors enhances the quality of predictive models. Such an integrated QSAR (I-QSAR) approach has been explored for predictive toxicology [Gute et al, *ETAP* 2004], and further I-QSAR modeling is planned using biodescriptors developed as part of this project.

5. Clustering and Molecular Similarity/ Dissimilarity Solutions

5.1 Chemical Clustering

Toxicity prediction models discussed above are convenient because they are useful in the straightforward prediction of the toxicity of chemicals from molecular structure or proteomics. But such an approach is difficult to use in cases of mixtures such as JP-8 primarily because it is a complex mixture of 230 (or 2,000?) distinct chemical ingredients. JP-8 produces various effects such as skin irritation, light-headedness, and immunosuppression. To determine which particular combination(s) of the 230! Or 2,000! possible candidates cause these effects is a daunting task.

One viable solution has been proposed by Basak and colleagues. In collaboration with other AFOSR grantees, the NRRI research team has devised methods for grouping JP-8 chemicals into a limited number of clusters so that a much smaller number of combinations of JP-8 chemicals can be tested in the lab. Clustering has been done using calculated TS and TC descriptors only because very limited property data are available for the JP8 chemicals.

5.2 Tailored Similarity

The research group's work to develop and refine quantitative molecular similarity analysis (QMSA) methods has also continued throughout this project. One recent addition to the field by our laboratory was the concept of tailored or property-specific QMSA methods [Basak, Gute et al, *SAR & QSAR* 2002; Basak et al, *THEOCHEM* 2002; Gute et al, *IEJMD* 2003; Gute and Basak, *SAR & QSAR* submitted]. Whereas ordinary similarity methods are based on an arbitrary set of descriptors or at best a set of parameters chosen subjectively based on previous experience of the scientist, the tailored QMSA method chooses a subset of indices strongly correlated with the property of interest and then uses those descriptors to created the *n*-dimensional structure space used to measure the structural similarity/ dissimilarity of molecules. Results have shown that tailored QMSA (T-QMSA) outperforms standard (or arbitrary) QMSA methods. It is expected that this novel methodology will find wide applications both in drug design and predictive toxicology and will lead to substantial technology transfer.

6. Statistical Methods for QSAR Model Validation

Development of useful quantitative structure-activity/ toxicity relationship (QSAR/QSTR) models has to be based on sound statistical methodology. Our group has been involved in the formulation of methods for the validation of QSAR models [Hawkins et al, *JCICS* 2003; Hawkins et al, *ETAP* 2004]. It is expected that such methods will be used widely in pharmaceutical and toxicological QSAR development. There were two major threads in the statistical work supported under the grant during the past project year.

6.1 Research into Optimization of Ridge Regression Techniques

One thread related to the use of linear modeling techniques in the high-dimensional underdetermined setting. This setting arises in the quantitative structure-xxx relationship (QSxR) modeling (xxx = property, activity, toxicity) which arises in several of the activities of the grant. Over the last few years, we have developed improved computational algorithms for ridge regression (RR) and implemented them in software which performs RR, partial least squares and principal component regression. This software has been applied to a number of QSxR problems and led to a number of publications that are reported elsewhere.

The software was initially seen as an algorithmically-enhanced implementation of a mature technology, but this has turned out not to be the case. Rather RR is well understood in the traditional setting with no more than a few dozen predictors, but still has some surprises in the very high dimensional QSxR work. This has led to some development in the statistical methodology area, and it seems it may become a focus of the PhD research of a student in the University of Minnesota School of Statistics.

6.2 Robust Statistical Methods for Large Data Arrays

A new and developing thread of work is in statistical methods for handling large data arrays that do not necessarily involve the prediction of a dependent. In cooperative work involving the National Institute for Statistical Sciences and two drug companies, Hawkins developed a robust method (rSVD) for performing principal component analysis in settings where missing data and/or outliers derail current standard statistical approaches. This technology had been proved on gene expression microarrays, as set out in more detail in a forthcoming PNAS paper, and a SELDI proteomics ovarian cancer data set.

This work is highly synergistic with that of the AFOSR grant since the setting of large arrays with some problematic data is also seen in two-dimensional electrophoresis gels such as are being developed by Dr. Witzmann and AFOSR collaborators. Initial application of the rSVD to 2D gels created following exposure of hepatocytes to various interesting organic chemicals has shown the method's ability to get past data problems but has not yet shown the expected connection with biological activity—presumably because the data seen so far have been confined to a small number of spots.

As this marriage of the problem and the potential tool is only a month old, much remains to be done to fine tune the method to the particular distinguishing features of 2D gels. However it does not seem premature to say that the methodology has the potential to help considerably in extracting information from the 2D gels that have until now presented problems of analysis.

V. Personnel Supported

Subhash C. Basak, Principal Investigator (NRRI)

Douglas Hawkins, Co-principal Investigator (Professor—Univ. of MN, St. Paul, MN)

Brian Gute, Research Fellow (NRRI)

Varsha Kodali, Graduate Research Assistant (NRRI)

Christian Mattson, Graduate Research Assistant (NRRI)

Denise Mills, Junior Scientist (NRRI)

Krishnan Balasubramanian, Consultant (Professor—Univ. of CA–Davis, Livermore, CA)

Marjan Vračko, Consultant (Professor—National Institute of Chemistry, Slovenia)

Milan Randić, Consultant (Distinguished Professor—Drake Univ., Des Moines, IA)

VI. Publications

The following 44 peer-reviewed papers, which are currently either published, in press, or submitted, report results of research carried out between March 1, 2002 and February 28, 2005.

2002

Alkane ordering as a criterion for similarity between topological indices: Index J as a "sharpened Wiener index", A.T. Balaban, D. Mills and S.C. Basak, *MATCH (Commun. Math. Comput. Chem.)*, **45**, 5–26 (2002).

A comparative study of proteomics maps using graph theoretical biodescriptors, M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, **42**, 983–992 (2002).

Novel matrix invariants for characterization of changes of proteomics maps, M. Randić, J. Zupan, M. Novič, B.D. Gute, and S.C. Basak, *SAR QSAR Environ. Res.*, **13**, 689–703 (2002).

Prediction of tissue-air partition coefficients: A comparison of structure-based and property-based methods, S.C. Basak, D. Mills, D.M. Hawkins, H.A. El-Masri, *SAR QSAR Environ. Res.*, **13**, 649–665 (2002).

QSAR modeling of flotation collectors using principal components extracted from topological indices, R. Natarajan, I. Nirdosh, S.C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.*, **42**, 1425–1430 (2002).

Quantitative descriptor for SNP related gene sequences, A. Nandy, P. Nandy and S.C. Basak, *Internet Electron. J. Mol. Des.*, **1**, 367–373 (2002), http://www.biochempress.com.

Quantitative molecular similarity analysis (QMSA) methods for property estimation: A comparison of property-based, arbitrary, and tailored similarity spaces, S.C. Basak, B.D. Gute, D. Mills, *SAR QSAR Environ. Res.*, **13**, 727–742 (2002).

Structure-water solubility modeling of aliphatic alcohols using the weighted path numbers, D. Amic, S.C. Basak, B. Lučić, S. Nikolić and N. Trinajstić, *SAR QSAR Environ. Res.*, **13**, 281–295 (2002).

Tailored similarity spaces for the prediction of physicochemical properties, B.D. Gute, S.C. Basak, D. Mills, and D.M. Hawkins, *Internet Electron. J. Mol. Des.*, **1**, 374–387 (2002), http://www.biochempress.com.

2003

Assessing model fit by cross-validation, D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.*, **43**, 579–586 (2003).

How and why did our view of the world change during the last six hundred years? A.T. Balaban and S.C. Basak, *Amer. Romanian Acad. J.*, **28**, 25-35 (2003).

Novel map descriptors for characterization of toxic effects in proteomics maps, Z. Bajzer, M. Randić, D. Plavšić, S.C. Basak, *J. Mol. Graph. Model.*, **22**, 1–9 (2003).

Predicting mutagenicity of congeneric and diverse sets of chemicals using computed molecular descriptors: A hierarchical approach, S.C. Basak, D. Mills, B.D. Gute and D.M. Hawkins, in *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, R. Benigni, Ed., CRC Press, Boca Raton, FL, Chapter 7, pp. 207–234.

Prediction of cellular toxicity of halocarbons from their computed chemodescriptors: A hierarchical QSAR approach, S.C. Basak, K. Balasubramanian, B.D. Gute, D. Mills, A.

Gorczynska and S. Roszak, *J. Chem. Inf. Comput. Sci.*, **43**, 1103–1109 (2003).

Prediction of human blood:air partition coefficient: A comparison of structure-based and property-based methods, S.C. Basak, D. Mills, D.M. Hawkins, and H. El-Masri, *Risk Analysis*, **23**, 1173-1184 (2003).

Quantitative molecular similarity methods in the property/ toxicity estimation of chemicals: A comparison of arbitrary versus tailored similarity spaces, S. C. Basak, B. D. Gute, D. Mills and D. M. Hawkins, *J. Mol. Struct.: THEOCHEM*, **622**, 127–145 (2003).

Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding potency of dibenzofurans: A hierarchical QSAR approach, S.C. Basak, D. Mills, M.M. Mumtaz and K. Balasubramanian, *Ind. J. Chem.*, **42A**, 1385–1391 (2003).

2004

Characterization of 2-D proteomics maps based on the nearest neighborhoods of spots, M. Randic, N. Lers, D. Plavsic and S.C. Basak, *Croat. Chem. Acta.*, **77**, 345–351 (2004).

Counter-propagation artificial neural network as a tool for the independent variable selection: Structure-mutagenicity study on aromatic amines, A. Jezierska, M. Vracko and S.C. Basak, *Mol. Diversity*, **8**, 371–377 (2004).

Interrelationship of major topological indices, S.C. Basak, A.T. Balaban and B.D. Gute, *Croat. Chem. Acta.*, **77**, 331–344 (2004).

Modeling of structure-mutagenicity relationship: counter propagation neural network approach using calculated structural descriptors, I.V. Valkova, M. Vracko and S.C. Basak, *Anal. Chim. Acta*, **509**, 179–186 (2004).

On invariants of a 2-D proteome map derived from neighborhood graphs, M. Randic, N. Lers, D. Plavsic and S.C. Basak, *J. Proteome Res.*, **3**, 778–785 (2004).

Predicting blood:air partition coefficients using theoretical molecular descriptors, S.C. Basak, D. Mills, H.A. El-Masri, M.M. Mumtaz and D.M. Hawkins, *Environ. Toxicol. Pharmacol.*, **16**, 45-55 (2004).

Prediction of halocarbon toxicity from structure: a hierarchical QSAR approach, B.D. Gute, K. Balasubramanian, K. Geiss and S.C. Basak, *Environ. Toxicol. Pharmacol.*, **16**, 121–129 (2004).

QSAR study using topological indices for inhibition of carbonic anhydrase II by sulfanilamides and Schiff bases, A.T. Balaban, S.C. Basak, A. Beteringhe, D. Mills and C.T. Supuran, *Mol. Diversity*, **8**, 401–412 (2004).

QSARs for chemical mutagens from structure: Ridge regression fitting and diagnostics, D.M. Hawkins, S.C. Basak, and D. Mills, *Environ. Toxicol. Pharmacol.*, **16**, 37–44 (2004).

Similarity study of proteomic maps, M. Vracko and S.C. Basak, *Chemometr. Intell. Lab. Syst.*, **70**, 33–38 (2004).

Structure-mutagenicity modeling using counter propagation neural networks, M. Vracko, D. Mills and S.C. Basak, *Environ. Toxicol. Pharmacol.*, **16**, 25–36 (2004).

Usefulness of graphical invariants in quantitative structure-activity correlations of tuberculostatic drugs of isonicotinic acid hydrazide type, M.C. Bagchi, B.C. Maiti, D. Mills, S.C. Basak, *J. Mol. Modeling*, **10**, 102–111 (2004).

Variable connectivity index as a tool for modeling structure-property relationships, M. Randic,

M. Pompe, D. Mills and S.C. Basak, *Molecules*, http://www.mdpi.org, 9, 1177–1193 (2004).

2005

Four-color map representation of DNA or RNA sequences and their numerical characterization, M. Randić, N. Lers, D. Plavšić, S. C. Basak, and A. T. Balaban, Chem. Phys. Lett., 407, 205–208 (2005).

In press

Canonical labeling of proteome maps, M. Randić, N. Lerš, D. Vukičević, D. Plavšić, B.D. Gute and S.C. Basak, *J. Proteome Res.*

On similarity of proteome maps, M. Randić and S. C. Basak, *Med. Chem. Res.*

Use of mathematical structural invariants in analyzing combinatorial libraries: A case study with Psoralen derivatives, S.C. Basak, D. Mills, B.D. Gute, A.T. Balaban, K. Basak and G.D. Grunwald, In *Some Aspects of Mathematical Chemistry*, Eds. D.K. Sinha, S.C. Basak, R.K. Mohanty and I.N. Basumallick, Visva-Bharati University: Santiniketan, West Bengal, India.

Use of proteomics based biodescriptors in the characterization of chemical toxicity, Z. Bajzer, S.C. Basak, M. Vracko Grobelsek and M. Randic, in *Genomic and Proteomic Applications of Toxicity Testing*, M.J. Cunningham, Ed., Humana Press, Inc.: Totowa, NJ.

Variable molecular descriptors, M. Randić and S.C. Basak, In *Some Aspects of Mathematical Chemistry*, Eds. D.K. Sinha, S.C. Basak, R.K. Mohanty and I.N. Basumallick, Visva-Bharati University: Santiniketan, West Bengal, India.

Submitted

Application of information-theoretic biodescriptors in toxicity prediction, S.C. Basak, B.D. Gute and F.A. Witzmann, *J. Chem. Inf. Model.*

Information-theoretic biodescriptors for proteomics maps: Development and applications in predictive toxicology, S.C. Basak, B.D. Gute and F.A. Witzmann, in *Proceedings of the 9th WSEAS International Conference on Computers.*

Modeling of aryl hydrocarbon (Ah) receptor binding affinity for dibenzofurans using the hierarchical quantitative structure-activity relationship (HiQSAR) approach, S.C. Basak and D. Mills, in the series *Chemistry of Biologically Active Synthetic and Natural Compounds*, special volume *Oxygen- and Sulfur-Containing Heterocycles, V.* Karstev, Ed., proceedings of the II International Conference on the Chemistry and Biological Activity of Oxygen- and Sulfur-Containing Heterocycles.

On the dependence of a characterization of proteomics maps on the number of protein spots considered, M. Randić, F.A. Witzmann, V. Kodali and S.C. Basak, *J. Chem. Inf. Model.*

Ordering of molecules using topological indices: Comparison of arbitrary and tailored structure spaces, B.D. Gute and S.C. Basak, *SAR QSAR Environ. Res.*

Predicting bioactivity and toxicity of chemicals from mathematical descriptors: A chemical-cum-biochemical approach, S.C. Basak, D. Mills and B.D. Gute, in *Advances in Quantum Chemistry: Chemical Graph Theory: wherefrom, wherefor, & whereto*, D.J. Klein and E. Brandas, Eds., Elsevier–Academic Press.

Proteomics 2-D map invariants based on the neighborhood graphs, M. Randić, D. Plavšić, and S.C. Basak, *J. Proteome Res.*

Quantitative structure-activity relationships (QSARs), S.C. Basak, D. Mills and B.D. Gute, in *Biological Concepts and Techniques in Toxicology: An Integrated Approach*, J. Riviere, Ed., Marcel-Dekker, Inc.

VII. Interactions/Transitions

a) Participation at Meetings

1. *Estimation of Toxicological and Ecotoxicological Properties of Chemicals from Structure: A Mathematical-cum-Computational Approach*, presented to the faculty and students of the Toxicology Program of the University of Minnesota, broadcast from Duluth, MN (March 2002).
2. *Numerical Graph Invariants: Development and Applications in Drug Discovery and Risk Assessment of Chemicals*, presented to the Department of Mathematics and Statistics, University of Minnesota Duluth, Duluth, MN (March 2002).
3. *Interrelationship of Major Topological Indices*, co-authored by B.D. Gute and A.T. Balaban, was presented at the Joint Annual Meeting of the Society of Environmental Toxicology and Chemistry (Midwest Chapter) and Society of Toxicology (Northland Chapter) organized at the USEPA Mid- Continental Ecology Division, in Duluth, MN (April 2002).
4. *Development of New Tools for Quantitative Characterization of Proteomics Maps*, at the symposium Development and "Application of Ecogenomics for Water Quality Assessment" organized jointly by the Council of State Governments and the United States Environmental Protection Agency, in Kansas City, MO (May 2002).
5. *Chemodescriptors Versus Biodescriptors in Toxicity Prediction of Halocarbons*, co-authored by B.D. Gute, D. Mills, K. Balasubramanian (UC - Davis), K. Geiss (Wright Patterson Air Force Base, Dayton, OH), D. Hawkins (U of MN, TC Campus), M. Randić (NRRI and National Institute of Chemistry, Ljubljana, Slovenia), F. Witzmann (Indiana University School of Medicine), and M. Vračko (National Institute of Chemistry, Ljubljana, Slovenia), was presented at the QSAR 2002 Conference, May 25-29, Ottawa, Canada.
6. *Predicting Toxicity of Chemicals in the Post-Genomic Era: A Computational Approach*, delivered to the Department of Molecular Biology and Genetics, at the University of Guelph, Guelph, Ontario, Canada (June 2002).
7. *Predicting Bioactivity of Chemicals from Structural and Proteomics-Based Descriptors*, presented at the University of Texas Medical Branch (UTMB) sponsored jointly by UTMB and Department of Marine Sciences, Texas A & M University at Galveston, Galveston, TX (June 2002).
8. Dr. Basak gave the following invited seminar lectures on predictive toxicology, QSAR and Mathematical Chemistry during his recent trip to India:
   i. *Predicting therapeutic activity, health hazard, and ecotoxicity of chemicals using structural and and proteomics based descriptors: An integrated approach*, July 26, 2002, at the Bose Institute, Kolkata, India;
   ii. A two part lecture on *Predicting bioactivity and toxicity of chemicals from structure and proteomics based descriptors* at the Department of Biochemistry, Calcutta

University, Kolkata, August 2, 2002;

- **iii.** A two part lecture on *Use of discrete mathematics in chemistry, drug design and ecotoxicology*, "Part I: Molecular descriptors and QSAR," August 7, 2002; "Part II: Molecular similarity and integrated QSAR in predictive toxicology," August 8, 2002, at the Sivatosh Mookerjee Science Center, Kolkata, India;
- **iv.** An invited lecture on *Integration of chemoinformatics and bioinformatics for drug discovery and environmental protection: An integrated approach in the post-genomic era*, at the Indian Institute of Chemical Biology, Kolkata;
- **v.** *Use of mathematical invariants in drug discovery and toxicology in the post-genomic era: An in silico approach.*

9. Subhash Basak was an invited speaker at the international conference *Thirty First Year of the Topological Index Z*, organized by Ochanomizu University (Tokyo) and National Institute of Advanced Industrial Science and Technology (Japan), October 28-29, 2002. He presented the following papers at the conference:

10. *Numerical graph invariants: Development and applications in drug discovery and risk assessment of chemicals* by Subhash C. Basak

11. *On Clustering of JP-8 Chemicals using graph invariants*, by Basak, Brian D. Gute, Gregory D. Grunwald, and James Riviere (College of Veterinary Medicine, North Carolina State University, Raleigh, NC)

12. *Quantitative molecular similarity methods in the estimation of chemical properties and activities: A comparison of arbitrary versus tailored similarity Spaces* by Gute, Basak, Denise Mills (all of NRRI) and Douglas M. Hawkins (School of Statistics, University of Minnesota, Minneapolis, MN)

13. Basak traveled to India to participate in a conference entitled *Science and Economy: New Insights on Medicinal Plants and Floriculture* organized by Darjeeling Doohrs Postgraduates' Welfare Association (DDPGWA), Darjeeling, West Bengal, India, and Institute of Rural Development, Natural Disaster and Environmental Management, Kolkata, India, November 17, 2002. Basak was the chief guest in the conference and gave a presentation entitled *Indigenous medicinal plants and modern drug discovery: A computational approach in the post-genomic era.*

14. Subhash Basak traveled to India, China, Croatia, and Slovenia during Jan 26 to March 5, 2003, to discuss collaborative research with colleagues and give the following invited lectures:

- **i.** *Use of Chemo and Bio-descriptors for predicting activity/toxicity of chemicals: An Integrated QSAR approach* at the University Department of Chemical Technology (UDCT), University of Bombay, Mumbai, India.
- **ii.** *Use of Chemodescriptors in drug design and hazard assessment of chemicals*, at the Center for Bioinformatics, University of Pune, Maharashtra, India.
- **iii.** *Integrated modeling of bioactivity using chemo and biodescriptors*, University of Pune.
- **iv.** *Mathematical invariants in the characterization of chemical structure, drug design, and environmental protection*, at the Thirty Second Annual Conference of the Association for the Improvement of Mathematics Teaching (AIMT), organized at the premises of Vivekananda Math, Ramakrishna Vivekananda Mission, Barrackpore, North 24 Parganas, West Bengal, India.

v. *Characterization of complex Ayurvedic medicinal products: A structural-cum-mathematical approach*, at the J. B. Ray Ayurvedic College, Kolkata, India.

vi. *Use of mathematical structural invariants to predict properties of molecules and biomolecules*, at the Department of Mathematics, University of Xiamen, Fujian, Peoples Republic of China.

vii. *Use of molecular descriptors in predicting pharmacological and toxicological properties of molecules*, Department of Chemistry, Xiamen University, P R China.

viii. *Integrated QSAR for drug design and natural medicine: A computational approach*, Xiamen University.

ix. *Drug Design using mathematical chemical descriptors*, at the S. N. Pradhan Center for Neurosciences, University College of Medicine, University of Calcutta, Kolkata, India.

x. *Chemodescriptors and biodescriptors: Development and applications in drug discovery and hazard assessment of chemicals*, at the National Institute of Chemistry, Ljubljana, Slovenia.

xi. *Applications of Chemodescriptors and biodescriptors in predicting therapeutic and toxic properties of chemicals*, at the Rugjer Boskovic Institute, The Republic of Croatia.

15. Basak and collaborators presented the following papers at the 2003 Toxicology and Risk Assessment Conference, April 28- May 1, in Fairborn, Ohio:

i. *Estimation of blood: Air partition coefficients of volatile organic chemicals using molecular descriptors*, authored jointly by Basak, Hisham El-Masri, Moiz M. Mumtaz (both of the Agency of Toxic Substances and Disease Registry, Centers for Disease Control and Prevention, Atlanta, GA), Douglas M. Hawkins (School of Statistics, U of M TC campus), Brian D. Gute and Denise Mills (both of NRRI).

ii. *Hierarchical quantitative structure-toxicity relationship (Hi-QSTR) modeling of chemical toxicity*, authored by Basak, Gute, and Mills.

iii. *A Comparison of arbitrary versus tailored similarity spaces in property/toxicity estimation*, authored by Gute, Basak, Mills and Hawkins.

iv. *Hierarchical QSAR analysis for the toxicity prediction of halocarbons*, by Gute, Basak, Krishnan Balasubramanian (University of California - Davis, Livermore, CA), Kevin Geiss (U.S. Air Force, Wright Patterson Air Force Base, OH); and Hawkins.

v. *Similarity studies of proteomic maps*, by Marjan Vracko (National Institute of Chemistry, Ljubljana, Slovenia) and Basak.

vi. *Characterization of toxicant-induced variations of proteomics maps*, by Milan Randic (NRRI and National Institute of Chemistry, Slovenia), Zelko Bajzer (Mayo Clinic, Rochester, MN), Basak, Gute (NRRI), and Dejan Plavsic (Institute Rudjer Boskovic, Zagreb, Croatia).

16. *Use of calculated structural descriptors in predicting toxicity of chemicals*, presented at the Conference on the Prediction of Acute Toxicity, May 1-2, organized by the MITRE Corporation, Washington, DC.

17. *Chemodescriptors and biodescriptors: Development and applications in drug design and hazard assessment of chemicals*, presented at Stanford University, Palo Alto, CA.

18. *Predicting membrane permeability using HiQSAR*, cooperatively authored by Basak, Jim Riviere (Center for Chemical Toxicology Research and Pharmacokinetics, College of Veterinary Medicine, North Carolina State University), Gute, Mills and Balasubramanian, presented at the JP8 Toxicology Conference, May 14-16, organized at the University of Arizona Medical Center, Tucson, AZ.

19. Session chair at the Chemoinformatics Symposium at the Intelligent Drug Discovery and Development Conference, May 28-30, organized by the Cambridge Health Institute, Philadelphia, Penn.

20. *The role of chemodescriptors and biodescriptors in drug design and predictive toxicology*, presented at the Chemoinformatics Symposium at the Intelligent Drug Discovery and Development Conference, May 28-30, organized by the Cambridge Health Institute, Philadelphia, Penn.

21. *Use of chemodescriptors and proteomics- based biodescriptors in predictive toxicology: A computational approach*, a two-part lecture presented at the ImageTox conference of the European Community organized at the University of Tartu, Estonia (June 2003).

22. *Mathematical descriptors in predicting activity/toxicity of chemicals*, presented at the Department of Crystallography, Birkbeck College, University of London (June 2003).

23. Session chair at the Sixth Girona Seminar on Molecular Similarity, July 21-24, organized at the University of Girona, Spain.

24. *Predicting toxicity of chemicals using chemodescriptors and biodescriptors*, presented at the Sixth Girona Seminar on Molecular Similarity, July 21-24, organized at the University of Girona, Spain.

25. Basak and collaborators presented the following papers at the Third Indo-US Workshop for Mathematical Chemistry, August 2-7 (2003), at the University of Minnesota Duluth, Duluth, MN, USA:

    i. *Applications of chemodescriptors and biodescriptors in predictive toxicology: An integrated approach*, authored by Basak.

    ii. *Self-organizing map and counter propagation neutral network as a tool in structure-property modeling*, authored jointly by Marjan Vracko, Aneta Jezierska, Iva Valkova, Paolo Mazzatorta, Emilio Benfenati, Basak, and Mills.

    iii. *Differential protein expression data derived from toxicoproteomics approaches*, jointly authored by Frank Witzmann, Heather Coppage, Junyu Li, Nathan Pedrick, Basak, Geiss, and John Frazier.

    iv. *Use of calculated molecular descriptors in quantitative structure activity relationship modeling of two-substituted isonicotinic acid hydrazide*, jointly authored by M.C. Bagchi, B.C. Maiti, Mills and Basak.

    v. *Hierarchical quantitative structure-toxicity relationship (Hi-QSTR) modeling of chemical toxicity*, jointly authored by Basak, Gute, and Mills.

    vi. *Estimation of tissue:air partition coefficients: A comparison of structure and property-based methods*, jointly authored by Basak, Mills, Hawkins, and El-Masri.

    vii. *Estimation of estrogen receptor binding affinity using theoretical molecular descriptors*, jointly authored by Basak, Mills, and Hawkins.

    viii. *Estimation of blood:air partition coefficients of volitile organic chemicals using molecular descriptors*, jointly authored by Basak, Gute, Mills, El-Masri, Mumtaz, and Hawkins.

    ix. *Use of topological indices in predicting aryl hydrocarbon (AH) receptor binding affinity of dibenzofurans: A hierarchical QSAR approach*, jointly authored by Basak, Mills, Mumtaz, and Balasubramanian.

x. *Heirarchical QSAR analysis for the toxicity prediction of halocarbons,* jointly authored by Gute, Basak, Balasubramanian, Geiss, and Hawkins.

xi. *Use of calculated chemodescriptors in toxicity prediction of halocarbons,* jointly authored by Gute, Basak, Balasubramanian, Geiss, and Hawkins.

xii. *A comparison of arbitrary vs. tailored similarity spaces in property/toxicity estimation,* jointly authored by Gute, Basak, Mills, and Hawkins.

26. Dr. Basak chaired a session and participated in a panel discussion at the International Symposium on Water: Crisis and Strategies, organized jointly by the International Institute of Bengal Basin and Center for Ground Water Studies in Kolkata, India, February 7–8, 2004.

27. Dr. Basak also gave an invited lecture entitled "Use of mathematical biodescriptors and chemodescriptors in predicting toxicity of chemicals" at the International Symposium on Water: Crisis and Strategies.

28. Dr. Basak made the following presentations at the Current Trends in Drug Discovery Research (CTDDR) conference in Lucknow, India, February 17–20, 2004:

i. An invited lecture "Prediction of bioactivity/ toxicity of chemicals using proteomics based mathematical descriptors" authored jointly by Basak and Brian Gute (NRRI), Marjan Vracko (National Institute of Chemistry, Ljubljana, Slovenia) and Milan Randic (NRRI and National Institute of Chemistry, Ljubljana, Slovenia)

ii. A paper titled "Applications of theoretical molecular descriptors in QSAR modeling of anti-mycobacterial compounds" authored jointly by Manish Bagchi and Bhim Maiti (Indian Institute of Chemical Biology, Kolkata, India), Denise Mills and Basak (NRRI)

29. Dr. Basak was a co-chairperson of the session "Toxicogenomics and Molecular Mechanisms" at the 11th International Workshop on Quantitative Structure-Activity Relationships in the Human Health and Environmental Sciences.

i. Dr. Basak gave an invited lecture entitled "Prediction of toxicity of chemicals using chemodescriptors and biodescriptors" at the workshop.

30. Basak and his collaborators from USA and Europe presented the following papers at the 11th International Workshop on Quantitative Structure-Activity Relationships in the Human Health and Environmental Sciences:

i. "Proteomics based biodescriptors for characterization of toxicity," authored by Basak, Milan Randic (National Institute of Chemistry, Ljubljana, Slovenia and NRRI), Dejan Plavsic (Institute Rudjer Boakovi, Zagreb, Croatia), and Željko Bajzer (Mayo Clinic, Rochester, MN)

ii. "A comparison of arbitrary versus tailored similarity spaces in property/toxicity estimation" by Brian Gute, Basak and Denise Mills (NRRI) and Douglas Hawkins (Univ. of MN)

iii. "Use of calculated chemodescriptors in toxicity prediction of halocarbons by Brian Gute and Basak (NRRI), Krishnan Balasubramanian (UC—Davis), Kevin Geiss (AFRL—WPAFB) and Douglas Hawkins (Univ. of MN)

iv. "Intercorrelation pattern of major topological indices" by Basak and Brian Gute (NRRI) and Alexandru Balaban (Texas A&M Univ.)

v. "On clustering of JP-8 chemicals" by Basak, Brian Gute and Gregory Grunwald (NRRI) and James Riviere (NCSU)

vi. "Estimation of tissue:air partition coefficients: a comparison of structure- and property-based methods" by Basak and Denise Mills (NRRI), Douglas Hawkins (Univ. of MN) and Hisham El-Masri (ATSDR)

vii. "Estimation of blood:air partition coefficients of volatile organic chemicals using molecular descriptors" by Basak, Brian Gute and Denise Mills (NRRI), Hisham El-Masri and Moiz Mumtaz (ATSDR) and Douglas M. Hawkins (Univ. of MN)

viii. "Estimation of estrogen receptor binding affinity using theoretical molecular descriptors" by Basak and Denise Mills (NRRI) and Douglas Hawkins (Univ. of MN)

ix. "Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding affinity of dibenzofurans: A hierarchical QSAR approach" by Basak and Denise Mills (NRRI), Moiz Mumtaz (ATSDR) and Krishnan Balasubramanian (UC—Davis)

x. "Similarity studies of proteomic maps" by Marjan Vracko (National Institute of Chemistry, Ljubljana, Slovenia), Basak, Kevin Geiss (AFRL—WPAFB) and Frank Witzmann (IUPUI)

31. Subhash Basak gave an invited lecture entitled "Use of mathematical biodescriptors and chemodescriptors in predicting bioactivity/toxicity of chemicals" at the Memorial Symposium on Molecular Informatics, Modeling and Simulation, on June 23–26, 2004 at Memorial University of Newfoundland, St John's, Newfoundland, Canada.

32. Subhash Basak and coworkers presented the following papers at the Southwest Theoretical Chemistry Conference XXI, Galveston, TX, October 22–23, 2004. Basak and coworkers also made the following presentations at the conference:

i. "Mathematical chemodescriptors and biodescriptors in biological structure-activity relationships" authored by Basak

ii. "Intercorrelation pattern of major topological Indices," by Basak, Brian D. Gute and Alexandru T. Balaban (Texas A&M Univ)

iii. "Prediction of blood:brain penetration of chemicals using computed molecular descriptors," by Christian T. Matson (NRRI), Lester R. Drewes (UMD School of Medicine) and Basak

iv. "Hierarchical quantitative structure-activity relationship (HiQSAR) studies of mosquito repellency of alicyclic carboxamides," by Basak, Ramanathan Natarajan (NRRI) and Denise Mills

v. "Predicting Antimycobacterial Activity of Quinolone Derivatives Using Theoretical Molecular Descriptors," by Mills, Manish C. Bagchi and Bhim C. Maiti (both of the Indian Inst of Chemical Biology) and Basak

vi. "Estimation of tissue:air partition coefficients: A comparison of structure- and property-based methods," by Basak and Mills, Douglas M. Hawkins (U of MN) and Hisham El-Masri (ATSDR)

vii. "Estimation of blood:air partition coefficients of volatile organic chemicals using molecular descriptors," by Basak, Gute, Mills, Hisham El-Masri and Moiz M. Mumtaz (ATSDR) and Hawkins (U of MN)

viii. "Estimation of estrogen receptor binding affinity using theoretical molecular descriptors," by Basak, Mills and Hawkins (U of MN)

ix. "Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding affinity of dibenzofurans: A hierarchical QSAR approach," by Basak, Mills, Moiz M. Mumtaz (ATSDR), Krishnan Balasubramanian (UC–Davis)

x. "Use of calculated chemodescriptors in toxicity prediction of halocarbons," by Gute, Basak, and Krishnan Balasubramanian (UC–Davis)

xi. "NMR spectral invariants as numerical descriptors for diastereomers," by Basak, Xiaofeng Guo and Fuji Zhang (both of Xiamen Univ, Peoples Republic of China) and Natarajan

xii. "Ordering the repellency of stereoisomers of topical mosquito repellents by molecular overlay," by Natarajan, Basak, Alexandru T. Balaban (Texas A&M Univ), Jerome A. Klun and Walter F. Schmidt (USDA)

xiii. "Stereochemical structure-activity relationship studies of insect repellents: A molecular mechanics approach," by Natarajan, Alexandru T. Balaban (Texas A&M Univ), Jerome A. Klun (USDA) and Basak

33. Basak and collaborators presented the following papers at the Fourth Indo-US Workshop for Mathematical Chemistry, organized in Pune, Maharashtra, India, January 8–12, 2005:

i. "Mathematical structural invariants: Developments and applications," by Subhash Basak

ii. "Partial order: A general tool in ecotoxicological and ecosystems hazard assessment," by Rainer Brüggemann (Inst of Freshwater Ecology and Inland Fisheries, Germany) and Basak

iii. "Prediction of biologic partition coefficients and binding affinities using SAR models," by Moiz M. Mumtaz and Hisham A. El-Masri (ATSDR), Douglas M. Hawkins (U of MN), Denise Mills and Basak

iv. "Quantitative structure-activity relationship (QSAR) modeling of juvenile hormone activity of alkyl (2E, 4E), 3, 7,11-trimethyl-2, 4-dodecadienoates," by Basak, Ramanathan Natarajan (NRRI), Mills and Brian Gute

v. "Tailored similarity: Creation of activity specific molecular similarity spaces," by Gute

vi. "On canonical labeling of proteins of proteomics maps," Milan Randic (Drake Univ.), Dejan Plavsic (Rugjer Boskovic Inst, Croatia) and Basak

vii. "Comparison of arbitrary versus tailored similarity spaces in property estimation," by Gute, Basak, Mills and Hawkins (U of MN)

viii. "Four-color map representation of DNA sequences and their numerical characterization," by Randic (Drake Univ), Basak, Nella Lera and Dejan Plavsic (both of Rugjer Boskovic Inst, Croatia) and Alexandru T. Balaban (Texas A&M Univ)

ix. "On invariants of a 2-D proteomic map derived from neighborhood graphs," by Randic (Drake Univ), Nella Lera and Dejan Plavsic (both of Rugjer Boskovic Inst, Croatia) and Basak

x. "Mutagen/non-mutagen classification of congeneric and diverse sets of chemicals using computed molecular descriptors: A hierarchical approach," by Basak, Mills, Gute, Christian Matson (NRRI) and Hawkins (U of MN)

xi. "NMR spectral invariants as numerical descriptors for diastereomers," by Basak and Natarajan

xii. "Ordering the repellency of stereoisomeric topical mosquito repellents by molecular overlay," by Natarajan, Basak, Alexandru T. Balaban (Texas A&M Univ), Jerome A. Klun and Walter F. Schmidt (both of USDA)

xiii. "Overall path connectivity—Two non-degenerate indices for alkanes," by Natarajan,

T. C. Murali and T. M. Anbazhagan (both of Crux Fusion Software, India)

- **xiv.** "Prediction of blood: brain penetration of chemicals using computed molecular descriptors," by Matson, Basak and Lester R. Drewes (UMD School of Medicine)
- **xv.** "QSTR models of juvenile hormone mimetic compounds for Culex pipiens larvae," by Jessica J. Kraker and Hawkins (both of the U of MN), Mills, Natarajan, and Basak
- **xvi.** "Similarity of proteomic maps: Use of similarity index and self-organizing maps," by Marjan G. Vracko (National Inst of Chemistry, Slovenia), Basak, Kevin Geiss (AFRL), Frank Witzmann (Purdue Univ of Indiana)
- **xvii.** "Similarity-based chemical clustering techniques," by Gute and Basak
- **xviii.** "Stereochemical structure-activity relationship studies of insect repellents," by Natarajan, Basak, Alexandru T. Balaban (Texas A&M Univ), and Jerome A. Klun (USDA)

34. Subhash Basak and colleagues presented the following papers at the conference Graph Theory and its Applications in Pollution Prevention and Drug Design: An International Conference Honoring Frank Harary, held on January 18, 2005 in Kolkata, India:

- **i.** "Advancing frontiers of graph-theoretic applications in environmental protection and drug design" an invited lecture by Basak
- **ii.** "Graph invariants and design of mosquito repellents," by Natarajan Ramanathan (NRRI) and Basak
- **iii.** "Prediction of blood:brain penetration of chemicals using computed molecular descriptors" by Christian T. Matson (NRRI), Lester R. Drewes (UMD School of Medicine) and Basak

35. Subhash Basak presented the paper "Computer-assisted estimation of blood:air partition coefficients of VOCs: Implications for chemical mixtures" authored jointly by Basak, Denise Mills (NRRI), Hisham A. El-Masri and Moiz M. Mumtaz (both of ATSDR) at the international conference Contemporary Concepts in Toxicology-Charting the Future: Building the Scientific Foundation for Mixtures Joint Toxicity and Risk Assessment, organized by the Agency for Toxic Substance and Disease Registry (ATSDR), Centers for Disease Control and Prevention, Atlanta, GA, February 16–17, 2005.

b) Consultative and Advisory Functions

1. Dr. Basak's research team has been consulting with Dr. Jim Riviere, another AFOSR grantee, and his colleagues Ronald Baynes and Xin-Rui Xia. Dr. Riviere's group is currently testing a selection of JP-8 constituents for skin penetration using a membrane coated fiber (MCF) method. Dr. Basak's team has consulted with them on the selection of test chemicals to better explore the JP-8 chemical structure space. An iterative process has been agreed upon, and the consultation will continue into the near future. Once results have been generated, the data will be used by Dr. Basak's team to develop models for dermal penetration.

2. Basak's research team has also continued its collaborative work with Dr. Jeff Fisher. Dr. Fisher's (AFRL) studies of JP-8 pharmacokinetics in animals will be used to develop computational models to predict the experimental data. Successful chemical structure based models will assist in the estimation of pharmacokinetics parameters for untested JP8 chemicals.

c) Transitions

1. Contribution to Computational Toxicology Program

The hierarchical QSAR approach for modeling toxicity (HiQSTR) developed by our research team has been adopted into the curriculum of the Masters coursework for the Predictive Toxicology program of the Environmental Sciences Department of the University of Calcutta, Kolkata, India. In addition, they have incorporated various molecular similarity methods, and just recently as a result of Dr. Basak's most recent visit to India, they have decided to add our recently developed tailored approach to molecular similarity (T-QMSA) into the curriculum as well.

2. Cooperative Work with Astra-Zeneca

We have recently begun a collaborative work with Dr. Indira Ghosh of Astra-Zeneca. We will be applying our newly developed tailored quantitative molecular similarity analysis (T-QMSA) techniques to a large proprietary drug discovery or pharmaceutical database provided by Astra-Zeneca.

VIII. New Discoveries, Inventions, or Patent Disclosures

For several years now, Dr. Basak's research team has pursued the concept of integrated quantitative structure activity relationship (I-QSAR) studies. This approach proposes that combining structural information about a chemical with biological response data should improve the ability to predict a biologically-relevant endpoint. At this stage, preliminary results from statistical modeling done both in Dr. Basak's laboratory and by Dr. Hawkins show that this is indeed the case—integrated models using both chemodescriptors and biodescriptors demonstrate an improved capacity to accurately model a biological response.

IX. Honors and Awards

a) Honors

1. 2004 American Statistics Association (ASA) Statistics in Chemistry Award

Drs. Basak and Hawkins, and Denise Mills were recognized by the ASA for their paper "Assessing Model Fit by Cross-Validation" appearing in the March-April 2003 issue of the *Journal of Chemical Information and Computer Sciences*, published by the American Chemical Society in Washington, D.C.

The Statistics in Chemistry Award recognizes outstanding collaborative endeavors between statisticians and chemists. Nominations are judged on two criteria: the innovative use of statistics to solve a problem in chemistry and the impact of the solution on the problem. The award was presented at the ASA Presidential Awards session in Toronto on August 10, 2004.

b) Advisory/Organizational Positions Held by Dr. Basak

1. Co-chair of the Indo-US Workshop in Mathematical Chemistry, an on-going biennial

conference series designed to promote an open exchange of scientific ideas between international scholars and with an emphasis on encouraging young Indian and American scientists (since 1997).

2. Member, editorial board of the international journal, *SAR and QSAR in Environmental Research* (Gordon and Breach).

3. Member, advisory board of the *Journal of Chemical Information and Computer Sciences* (American Chemical Society) (since 2003).

4. President of the International Society for Mathematical Chemistry (2003-2007).