Molecular diversity classification *via* information theory: A review

F. Torrens[1,*] and G. Castellano[2]

[1]Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, E-46071 València, Spain
[2]Cátedra Energesis de Tecnología Interdisciplinar, Universidad Católica de Valencia *San Vicente Mártir*, Guillem de Castro-94, E-46001 València, Spain

Abstract

Complexity definitions introduce context dependence. Molecular diversity is reviewed focusing on information theory. Compound set is viewed as static microstate collection registering information about environment. Method tends to oversample remote feature space and produce unbalanced designs. Results show limitation and provide rationale for failure. Affinity includes traverse ease via chreodes to effector. Lag results because of time/concentration needed for drug to displace transmitter molecules from chreodes. Molecules unfit for system are excluded from effector not fitting chreode patterns. Enzyme catalytic products leave active site at faster rate via chreodes minimizing delay of diffusion-controlled rate limitation. Interdisciplinary research in systems biology promises insights into life, evolution and disease organizational principles. Complex disorders represent umbrella terms for collections of conditions caused by rare, recent mutations in any of large number of genes.

1. Introduction

Systems are things that one is interested in, defined by boundaries and made up of *agents* [1]. Complexity is some-systems property, special kind of organization and associated with life. Systems are ranked: simple, complicated and complex. The former are understood from agents, present little change and describe most inanimate things; their properties are agent sum; they show little diversity; usually they are man-made and each is fungible. Complicated systems are learned from agents; functions are not related to parts; one lacks tools to learn. They are inanimate systems; numerical aspects are reoccurring events. Complex systems are not understood from agents, formed by different agents, always in motion, perform different functions and have history. Complexity results way to approach living systems: old method is reductionism; new

one results synergy. Approach is *systems biology*. Complex system characteristics are: agents are dynamic, type diversity is enormous, they self-organize and exist within hierarchy. Agents change; each has history; they interact locally with neighbours. No agent types and perpetual novelty exist. Their phenomena are related to self-organization: everything is dynamic and things aggregate. No orderliness master control exists, which is local. Diversity is everywhere; agent types/combinations exist where agents respond to crises, adapting to environmental changes; it exists in hierarchy. Each agent is complex system; each system exerts control over agents and is part of next higher order. Studies cannot exceed immediate relation called logical depth. Complex-system functions are: agents interact extensively, constant change in agent structure, position/properties and adaptation; diversity permits degree of interactions. In emergence, interaction changes agents producing new properties, which are not predictable from old ones; whole is more important than parts. Ways to study complex systems are to understand functions, synergy, create models and systems bioapproach.

---

* Corresponding author, Email: francisco.torrens@uv.es

Complexity was reviewed from trendy buzzword to emerging science [2]. It/emergence were discussed [3].

In earlier publications, fractal hybrid-orbital analyses of protein tertiary structure were performed [4]. Complexity, emergence and molecular diversity *via* information theory were analyzed. Valence topological charge-transfer indices for dipoles were obtained [5] and extended to homo/heterocycles and proteins [6]. Information-entropy molecular classification was applied to local anaesthetics [7,8] and inhibitors of human immunodeficiency virus type 1 (HIV-1) [9,10]. It was reported structural classification of complex molecules by artificial intelligence, information entropy and equipartition conjecture, *e.g.*, anti-cancer [11], phenolics [12], flavonoids, analgesics and cardiovascular system drugs [13]. It was calculated bond-based linear indices of non/stochastic edge adjacency matrices of organic physicochemical properties [14] and novel coumarin-based tyrosinase inhibitors discovered by quantitative structure–activity relationship (QSAR) validated by Organisation for Economic Co-operation and Development (OECD) principles [15]. The present report reviews molecular diversity classification based on information theory. The following section describes computational method. In the next section calculation results are presented/discussed. The final section summarizes our perspectives.

2. Computational method

Lin [16] assessed molecular diversity based on information theory [17–20]. Compound set is static molecular collection of microstates, which can register information about environment at predetermined capacity. Molecular diversity is related to population information content I as:

$$I = S_{max} - S \tag{1}$$

where S is system *entropy* given by von Neumann-Shannon expression:

$$S = -\sum_{i=1}^{n} p_i \ln p_i \tag{2}$$

where n is total number of microstates in system and p_i, *i-th* microstate probability subject to $\sum_{i=1}^{n} p_i = 1$. Each compound set represents finite number of distinguishable molecular species. System entropy results:

$$S(m,n) = -\sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} \ln p_{ij} \tag{3}$$

where m is number of species, n, that of individuals in population, and p_{ij}, probability of finding individual *i-th* in species *j-th*, which must satisfy $\sum_{j=1}^{n} \sum_{i=1}^{m} p_{ij} = 1$ in which maximum entropy [Equation (1)] is given by $S_{max}(m,n) = -n\ln m$. Collection information content rises as species

number decays. Difficulty stems because m is unknown. Each population member is unique, distinguishable species, and system entropy is related to species *distinguishability* rather than similarity to *a priori* known prototype set. Equation (3) is replaced by:

$$S(n,n) = -\sum_{j=1}^{n} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{4}$$

subject to $\Sigma\Sigma p_{ij} = 1$ and S_{max} becomes $S_{max}(n,n) = -n\ln n$. In Equation (4), p_{ij} are computed from molecular similarity table. Methods for quantifying similarity assign scores in [0,1], which involve computing similarity ρ_{ij} using established approach and normalization factor to derive probabilities [21,22]. Factor results $c = 1/\sum_{j=1}^{n}\sum_{i=1}^{n}\rho_{ij}$, and actual probabilities, $p_{ij} = c\rho_{ij}$. As probabilities close, species become less indistinguishable, system entropy decays and information registered by population rises.

3. Results and discussion

While information theory to quantify molecular diversity presents intellectual appeal, it results limited [23–25]. Figure (1) displays two sets of three imaginary compounds plotted *vs.* uniform properly scale. Distances between pairs are $d_{13} = d_{23} = 0.5$ and $d_{12} = 1.0$. Distance d_{ij} is taken as similarity measure ρ_{ij}. Two methods result:

$$\rho_{ij} = \alpha - d_{ij} \tag{5}$$

and:

$$\rho_{ij} = \frac{1}{1 + \alpha d_{ij}} \tag{6}$$

where α is constant (=1.0). Using Equation (4), one computes *entropies*, and sets $S_a = 1.887$ and $S_b = 1.609$ using linear form [Equation (5)], and $S_a = 2.163$ and $S_b = 2.144$ using reciprocal one [Equation (6)], where a and b refer to collections on left/right sides. According to Equation (1), set-a diversity is greater than b regardless of similarity functional form.

Figure 1. Two sets of three imaginary compounds *vs.* uniform properly scale; in (b) points 2 and 3 coincide

Figure (2) represents entropy of three-point set above *vs.* position of one point relative to the others. Both extremes denote cases where third point coincides with one of the two

reference points, while middle indicates situation depicted in Figure (1a). Profile is based on reciprocal function [Equation (6)] but similar results are obtained with Equation (5). Entropy function is at maximum when third point is located halfway, between two reference points, and monotonically decays as point moves from centre in either direction.

Symmetry decay decreases entropy. Diudea [26] edited a special issue on symmetry in nanostructures in *Symmetry: Culture and Science*. Entropy was extended [27–31], with illustrations in high-energy experiments in LHC-CERN/RHIC-Brookhaven and financial events with risk [32].

Table 1. Definitions of network and node parameters

Parameter	Type	Description
Degree	Node	Number of interactions
Average degree	Network	Average number of interactions for all nodes in a network
Power law coefficient	Network	Exponent in equation describing network degree distribution
Closeness centrality	Node	Reciprocal of sum of all shortest paths between a particular node and all other network nodes
Average closeness centrality	Network	Closeness centrality averaged for all network nodes
Average clustering coefficient	Network	Connectivity of all immediate neighbours of a particular node averaged for all network nodes
Average shortest path	Network	Average of all shortest paths (in number of edges) between all pairs of nodes in network
Density	Network	Ratio of actual number of edges in a network to possible number of edges
Diameter	Network	Longest shortest path across network

3.1. Network parameters/structure analysis

Network parameters are calculated to analyze net structure and identify key nodes. Definitions of network and node parameters used for net analyses are listed in Table 1 [33–37]. Atomic proximity referred to the other atoms is an index based on computer networks.
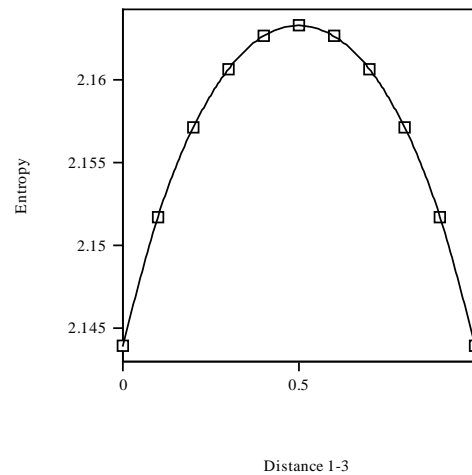

Distance 1-3

Figure 2. Variation of entropy as a function of *x*

3.2. Category theory and periodic table

Polystochastic model developments require running computations for processes of processes, *etc*. [38]. Stabilization hypothesis helps for difficult task [39,40]; it refers to k-tuply monoidal n-categories in which objects are multiplied in k ways, all of which interchange with each other up to isomorphism, which implies that k ways end up being equivalent but that single resulting operation is increasingly commutative as k rises. *Stabilization hypothesis* states that when one reaches $k = n + 2$, multiplication becomes maximally commutative; each column in n-category periodic table (PT) stabilizes at certain precise point. The PT for classifying n-categories contains $(n + k)$-category conjectured description with one j-morphism for $j < k$. Idea is to study n-category degenerate forms: n-categories that are trivial below certain dimension k. Such an n-category presents only non-trivial cells in top $n–k$ dimensions, so one performs dimension shift regarding this as $(n – k)$-category. Previous k-cells become new 0-cells, previous $(k + 1)$-cells undergo new 1-cells and previous n-cells turn into new $(n – k)$-cells, which is called k-fold degenerate n-category. The PT shows that $(n + k)$-category with one j-morphism for $j < k$ is reinterpreted as n-category; however, it will be n-category with k ways to multiply: k-tuply monoidal n-category (if $n = 1$, $k = 1$, 2-category with one object is monoidal category). The PT outlines properties (monoidal, braided, sylleptic, involutory and symmetric). In first row, $k = 0$, 0-monoidal n-category is n-category. In next row, $k = 1$, 1-monoidal n-category is monoidal n-category [1-monoidal 0-category is one-object category (monoid) and 1-monoidal 1-category is one-object 2-category (monoidal category)]. Monoidal 2-category is defined as one-object 3-category or

directly as 2-category with tensor. In third row, $k = 2$, degenerate monoidal category is commutative monoid and doubly degenerate 3-category is braided monoidal category. In first column, $n = 0$, one-object braided monoidal category is commutative monoid together with extra data for braiding satisfying axioms, which gives entry for $k = 3$, $n = 0$, and the same applies all way down column rest. Similar results are established for second column $n = 1$. For $k \geq 3$, k-monoidal 1-category is just symmetric monoidal category. Column stabilizes and stabilization point is most symmetric possible object. Sylleptic characterization is completed by more terms (involutory). Braided category is monoidal category with additional structure, sylleptic category is braided category with additional structure, *etc*. [41]. Table 2 lists PT of local anaesthetics, ice and benzyl alcohol..

It was reported reaction-net partially ordered set (poset) combinatorics [42], net-QSARs with reaction poset quantitative superstructure–activity relationships (QSSARs) for polychlorinated biphenyl (PCB) chromatography [43] and posets/substances' property ranking/prediction [44].

Table 2. Periodic table for local anaesthetics, ice and benzyl alcohol

g000	g010	g100	g101	g110	g111
ice		benzyl alcohol		diperodon pramoxine mexiletine	cocaine cyclomethycaine
	dibucaine propanolol		benzocaine butamben	dyclonine	hexylcaine piperocaine
	dimethisoquin			bupivacaine etidocaine lidocaine mepivacaine prilocaine tocainide (*S*)-ropivacaine	benoxinate proparacaine propoxycaine
		phenytoin			butacaine 2-chloroprocaine procaine tetracaine

3.3. Ligand diffusion over protein surfaces

Diffusion *via* bulk H_2O compares *vs*. across effector landscape; proximal relation exists among metabolic events; catalysis rate is biased directed diffusion; microviscosity is important. Smoluchowski introduced [45,46] and Debye enriched [47] diffusion-controlled reactions; three-dimensional (3D) random walk of ligand *via* bulk H_2O to effector is slow to accomplish effector complex tasks. Welch, Adam and Eigen agreed that two-dimensional (2D) surface diffusion to active site enhances diffusion-controlled reaction rate. Hasinoff concluded that fast enzyme reactions are limited by ligand diffusion to effector. Berg and Purcell described ligand skipping over cell surface with encountering-effector chance; cell membrane surface supports 2D diffusion but contains no variety to influence direction [48]. Rhodes, Sarmiento and Herbette invoked protein as target followed by lateral diffusion to effector across/*via* membrane. Blum described ligand guidance to effector *via* folds created by Bonnet transformation. Chen and Zhou explained that effector-outside protein is *promoter* causing ligand flow to effector. Sweet-tasting molecule administration produced response; washout created persistence. Birch postulated queue formation on receptor protein surface; exiting from tailback after washout provided persistence until backlog emptied; concept invoked washout-resistant residence on protein surface, favoured location and directional influence. H_2O around hydrophobic solute presents greater attraction for itself than solute; greater bent exists for H_2O to H-bond to nearby H_2O than hydrate solute, which molecules may/not bond depending on electrostriction: hydrophobic solutes occupy cavities while hydrophilic ones hydrate. H_2O on protein surface is *ordered* because of hydrophobic influences from side chains; ordering extends from protein surface, H_2O layers; H^+ exchange between two H_2O occurs at 10^{-14}s: landscape is hydrodynamic. Kier *et al*. built ligand diffusion models over protein surfaces. Waddington analyzed chreode (probability pathway) on epigenetic landscape. Enzyme reactions are fast because substrates diffuse to receptor and products diffuse from active site rapidly *via* chreodes [49,50]. H_2O molecules/nanostructures were reported [51].

3.4. *Grain* or *particle–pellet* model

Mathematical models were developed to describe progress of successive gas–solid reactions taking place in porous pellet, *e.g*., direct $H_{2(g)}$ reduction of $MoS_{2(s)}$ in $CaO_{(s)}$ presence [52–57]. Complexity arising from multi-step reactions causes relative deviation between model prediction and experiment. Complex-reaction modelling should include effects of solid structural change and heat balance (non-isothermal condition).

3.5. What is complex in complex disorders?

Compared to new chemical entities (NCEs), biologics are sensitive to stability, because of not only large molecular weight/size but also complex 3D structure that determines bioactivity: structural changes influence bioactivity [58]. In addition to chemical instabilities, *e.g.*, de-amidation, oxidation, isomerization of certain amino acids (AAs), hydrolysis, disulphide scrambling or glycation, physical/colloidal instabilities, *e.g.*, protein denaturation or formation of in/extrinsic protein particles, are relevant because of protein-aggregate immunogenicity. Process impacts on protein quality/integrity. Hlavacek *et al.* reviewed rules for modelling signal-transduction systems, proposing combinatorial complexity: number of possible protein complexes and protein-modification combinations increase exponentially [59]. Dopazo reported case of high-performance computing (HPC) in genomics, which analysis
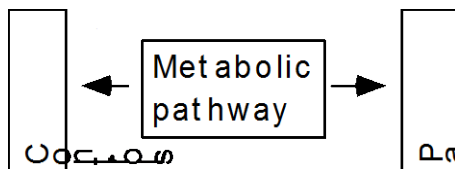

Figure 3. Finding mutations causative of diseases

requires extreme calculating to know, *e.g.*, which genome regions are methylated (switched off) [60]. Finding disease-causing mutations is complex problem (*cf.* Figure 3).

Principle is patient/family comparison to reference controls (*cf.* Figure 4); *e.g.*, how to find mutations associated with diseases? Illness is coded in genome: in monotonic diseases only one mutation causes illness; bioinformatics challenge is finding disease-causing mutation. Technical limit exists in analyzing genomes: <200 letters at a time; one must assemble data. Mutations change word meaning of genetic message. Challenge is to analyze mutated code fragments. In pipeline resequencing, problem is how to speed up pipe. Sequence data property is that they can be dealt in parallel. In ribonucleic acid (RNA), sequencing complexity of mapping transcripts exists: problem is finding matching regions. In deoxyribonucleic acid (DNA) *de novo* assembly, difficulty is that no genome reference is available.

Sequencing genomes of first eukaryotes seemed that gene number shows no correlation with organism complexity (paradox G-value); attempts tried to resolve inconsistency, *e.g.*, protein multifunctionality, alternative splicing, microRNAs or non-coding DNA. As intrinsic protein disorder was linked with complex responses to environmental stimuli/communication between cells, perhaps structural disorder increase species complexity. Schad *et al.* revisited paradox G-value, analyzing proteomes/complexity *via* number of cell types: they found that complexity/proteome size, measured by number of AAs, correlate showing power function; they analyzed complexity features in organisms/tissues finding: (1)

fraction of protein structural disorder increases between pro/eukaryotes but not over evolution, (2) number of predicted binding sites in disordered regions in proteome augments with complexity and (3) protein-disorder fraction, predicted binding sites, alternative splicing and protein/protein interactions rise with human tissue complexity [61]. Neurodevelopmental disorder encompasses disease range, *e.g.*, syndromes caused by rare mutations in specific genes/chromosomal loci, and more common disorders, *e.g.*, schizophrenia (SZ), autism spectrum and
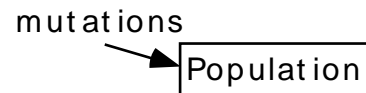
Figure. 4. Comparison of patients to controls.

idiopathic epilepsy/mental retardation [62]. Unravelling common-disorder genetics led to paradigm shift in genetic architecture of common neurodevelopmental disease, highlighting individual, rare mutations and overlapping genetic etiology. They converged on neurodevelopmental pathways, providing insights into pathogenic mechanisms. Paradigm stated that in individual with SZ, genetic risk is because of combination of genetic variants of small effect. Experiment prompted re-evaluation of phylogenic, common disease-common variant (CDCV) model. Evidence includes lack of expected positive findings from genome-wide association studies (GWASs) and discovery of mutations that predispose to SZ/psychiatric disorders, which led to mixed model wherein some cases are caused by polygenic mechanisms and others by single mutations. Model runs counter to theoretical literature body that supposed rejected Mendelian inheritance with genetic heterogeneity. Mitchell and Porteous asked how theory/data discrepancy arose and proposed evidence base rationalization [63]. They reconsidered theoretical analysis methods/conclusions and assumptions; they showed that conjectures are false, and genetic heterogeneity model is consistent with observed familial recurrence risks, endophenotype studies and population-wide parameters. Rather than polygenic, complex disorders represent umbrella terms for collections of conditions caused by rare, recent mutations in any of large number of genes [64]. Genetic-mutation list associated with human disease is long and grows and, with it, challenges of deciphering what goes wrong in sick people cells, tissues and bodies [65]. Problem is that ge/phenotype relation in health/disease is complex: it is exception when single mutation in single gene causes illness.

4. Perspectives

From results and discussion, perspectives of review follow.

1. Attributing complexity to only models and not to natural systems, and relativizing conception to chosen framework, one arrives at analytically useful idea of

complexity, which not only captures concept but also compares formulations illuminating philosophical problems.

2. We focused on particular complexity measure to capture decomposing-expression difficulty and sketched possible formal structure to relate different formal languages to complexity source analysis, which provided framework for systematic method simplification; work could form part of *complexity science*. More can be done: (a) research into syntactic structures/expression complexity relation, (b) development of semi/automatic simplification methods, (c) integration into work in *complex systems*, especially identifying possible complexity *causes*, (d) complexity-measure development and (e) use as complexity formulation synthesis. Way to extend work is towards model of representing process itself, which involves formalizing semantic picture and is seen as measure-theory extension to nonnumeric structures. As Badii and Politi put it in their book: *The natural extension of the study of complexity... seems, therefore, to point inevitably to a theory of model inference*.

3. Self-organizing modelling extracts hidden knowledge from data, serving for decision support of real-world problems; it is alternative to statistics, *neural networks* or *neurofuzzy* methods since it creates optimal complex models automatic, fast and systematically, and provides explanation component *via* explicit visible model descriptions.

4. Report described experience with approach assessing molecular diversity based on information theory. As demonstrated with examples, method shows trend to oversample remote areas of feature space producing unbalanced designs, which is because of certain information type whose mathematical definition is inappropriate for molecular diversity.

5. Structure-based subcellular pharmacokinetic models chemical behaviour/effects in biosystems, determined by chemical/biosystem physicochemical properties/structures. Absorption, distribution, metabolism, excretion and toxicity are predicted using empirical modelling with molecular descriptors using statistics, without attending mechanisms. Approaches capture pharmacokinetic complexity *via* unjustified descriptor use, rather than focusing on chemicobiological interactions and building mechanistic models, which are nonlinear in optimized coefficients.

6. Till not many time ago, implicit idea was that simple systems behave simply, and complex behaviour was complex-cause result; however, after chaos theory, the former systems can produce complex behaviours and the latter systems not necessarily carry associated complex answers. Knowledge removes control.

7. Topical anaesthetics remain powerful advancement for minimizing pain during cutaneous procedures. While topical agents were released with increased efficacy and faster onset, EMLA® (lidocaine/prilocaine 2.5/2.5 wt.%) remains most widely used one, given efficacy/safety proven by clinical trials. As practitioner options grow, comparison of onset of action, efficacy and safety is important. Our program MolClas classifies local anaesthetics/mixtures for difficult cases that are *a priori* hard to sort, *e.g.*, relation between procaine, ice, alcohols, *etc*. Ice/EMLA decay discomfort associated with needle injection: although EMLA performs better in pain control, ice shows advantages in easy of use, fast action and is lesser expensive; ice/EMLA are good agents each with dis/advantages in clinic. Benzyl alcohol is efficient anaesthetic for intact mucous membranes, surpassing procaine, ranking with alypine/β-eucaine and weaker than holocaine/cocaine; action is not as lasting as cocaine and 1% solutions produce considerable smarting; although anaesthesia duration provided by benzyl alcohol 0.9% is limited, advantages as local anaesthetic in minor plastic surgery include inexpensiveness/lesser adverse reactions; bacteriostatic saline preinjection decreases pain incidence/severity associated with propofol intravenous injection; decreased incidence is comparable to mixing lidocaine/propofol and makes alternative; however, use is not recommended in neonatal/pediatrics.

8. Complexity is multi-parametric trait determined by interaction potential, alternative splicing capacity, tissue-specific protein disorder and proteome size. Paradox G-value is apparent when plants are grouped with metazoans, because of different complexity/proteome size relation. It was argued for mixed model more bioconsilient, which involves disease-causing/modifying variant interactions in individual. Model implications were considered for moving schizophrenia beyond statistical associations to pathogenic mechanisms. Rather than polygenic, complex disorders represent umbrella terms for collections of conditions caused by rare, recent mutations in any of large number of genes.

Acknowledgements

References

[1] KIER, L. B., SEYBOLD, P. G. and CHENG, C. K. (2005) *Modeling Chemical Systems Using Cellular Automata* (Dordrecht: Springer).

[2] CASTI, C. (1992) The simply complex: Trendy buzzword or emerging new science? *Bull. Santa Fe Inst.* **7**:10-13.

[3] STANDISH, R. K. (2001) On complexity and emergence. *Complexity Int.* **9**:standi-1-6.

[4] TORRENS, F. (2000) Fractal hybrid orbitals in biopolymer chains. *Russ. J. Phys. Chem. (Engl. Transl.)* **74**:115-120.

[5] TORRENS, F. (2004) Valence topological charge-transfer indices for dipole moments. *Mol. Diversity* **8**:365-370.

[6] TORRENS, F. and CASTELLANO, G. (2009) Topological charge-transfer indices: From small molecules to proteins. *Curr. Proteomics* **6**:204-213.

[7] TORRENS, F. and CASTELLANO, G. (2006) Periodic classification of local anaesthetics (procaine analogues). *Int. J. Mol. Sci.* **7**:12-34.

[8] TORRENS, F. and CASTELLANO, G. (2011) Using chemical structural indicators for periodic classification of local anaesthetics. *Int. J. Chemoinf. Chem. Eng.* **1**(2):15-35.

[9] Torrens, F. and Castellano, G. (2010) Table of periodic properties of human immunodeficiency virus inhibitors. *Int. J. Comput. Intelligence Bioinf. Syst. Biol*. **1**:246-273.

[10] Torrens, F. and Castellano, G. (2011) Molecular classification of thiocarbamates with cytoprotection activity against human immunodeficiency virus. *Int. J. Chem. Model*. **3**:269-296.

[11] Torrens, F. and Castellano, G. (in press) Molecular classification by information theoretic entropy: Oxadiazolamines as potential therapeutic agents. *Curr. Comput.-Aided Drug Des*.

[12] Castellano, G., Tena, J. and Torrens, F. (2012) Classification of polyphenolic compounds by chemical structural indicators and its relation to antioxidant properties of *Posidonia oceanica* (L.) Delile. *MATCH Commun. Math. Comput. Chem*. **67**:231-250.

[13] Torrens, F. and Castellano, G. (submitted for publicayion) Molecular classification of cardiovascular system drugs (nifedypine analogues). *Curr. Phys. Chem*.

[14] Marrero-Ponce, Y., Martínez-Albelo, E. R., Casañola-Martín, G. M., Castillo-Garit, J. A., Echevería-Díaz, Y., Romero Zaldívar, V. Tygat, J., Rodriguez Borges, J. E., García-Domenech, R., Torrens, F. and Pérez-Giménez, F. (2010) Bond-based linear indices of the non-stochastic and stochastic edge-adjacency matrix. 1. Theory and modeling of *ChemPhys* properties of organic molecules. *Mol Divers*. **14**:731-753.

[15] Le-Thi-Thu, H., Casañola-Martín, G. M., Marrero-Ponce, Y., Rescigno, A., Saso, L., Parmar, V. S., Torrens, F. and Abad, C. (2011) Novel coumarin-based tyrosinase inhibitors discovered by OECD principles-validated QSAR approach from an enlarged balanced database. *Mol. Divers*. **15**:507-520.

[16] Lin, S. K. (1996) Molecular diversity assessment: Logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy of mixing. *Molecules* **1**:57-67.

[17] Shannon, C. E. (1948) A mathematical theory of communication: Part I, discrete noiseless systems. *Bell Syst. Tech. J*. **27**:379-423.

[18] Shannon, C. E. (1948) A mathematical theory of communication: Part II, the discrete channel with noise. *Bell Syst. Tech. J*. **27**:623-656.

[19] Brillouin, L. (1962) *Science and Information Theory* (London: Academic).

[20] Shannon, C. E. and Wheaver, W. (1963) *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois).

[21] Johnson, M. A. and Maggiora, G. M. [eds.] (1990) *Concepts and Applications of Molecular Similarity* (New York: Wiley-Interscience).

[22] Kubinyi, H. (1993) *QSAR: Hansch Analysis and Related Approaches* (Weinheim, Ger.: VCH), p. 172.

[23] Agrafiotis, D. K. (1997) On the use of information theory for assessing molecular diversity. *J. Chem. Inf. Comput. Sci*. **37**:576-580.

[24] Agrafiotis, D. K. (1997) Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci*. **37**:841-851.

[25] Agrafiotis, D. K. (1998) *Molecular diversity. In Encyclopedia of computational chemistry* (New York, NY: Wiley).

[26] Diudea, M. V. (2008) Symmetry in nanostructures. *Symmetry Cult. Sci*. **19**:229-231.

[27] Tsallis, C. (2009) *Introduction to Nonextensive Statistical Mechanics – Approaching a Complex World* (New York, NY: Springer).

[28] Tsallis, C. (2009) Entropy. In *Encyclopedia of Complexity and Systems Science* (Berlin: Springer).

[29] Umarov, S., Tsallis, C., Gell-Mann, M. and Steinberg, S. (2010) Generalization of symmetric α-stable Lévy distributions for $q > 1$. *J. Math. Phys*. **51**:033502–1-23.

[30] Andrade, Jr., J. S., da Silva, G. F. T., Moreira, A. A., Nobre, F. D. and Curado, E. M. F. (2010) Thermostatistics of overdamped motion of interacting particles. *Phys. Rev. Lett*. **105**:260601–1-4.

[31] Nobre, F. D., Rego-Monteiro, M. A. and Tsallis, C. (2011) Numerical relativistic and quantum equations with a common type of solution. *Phys. Rev. Lett*. **106**:140601–1-4.

[32] Ludescher, J., Tsallis, C. and Bunde, A. (2011) Universal behaviour of interoccurrence times between losses in financial markets: An analytical description. *Eur. Phys. Lett*. **95**:68002–1-5.

[33] Barabasi, A. L. and Oltvai, Z. N. (2004) Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet*. **5**:101-115.

[34] Zhu, X. W., Gerstein, M. and Snyder, M. (2007) Getting connected: Analysis and principles of biological networks. *Genes Dev*. **21**:1010-1024.

[35] Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Res*. **18**:644-652.

[36] Lin, W. H., Liu, W. C. and Hwang, M. J. (2009) Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein–protein interaction networks. *BMC Syst. Biol*. **3**.

[37] Hallock, P. and Thomas, M. A. (2012) Integrating the Alzheimer's disease proteome and transcriptome: A comprehensive network model of a complex disease. *OMICS* **16**:37-49.

[38] Iordache, O. (2011) *Modeling Multi-Level Systems* (Berlin: Springer).

[39] Baez, J. and Dolan, J. (1995) Higher dimensional algebra and topological quantum field theory. *J. Math. Phys*. **36**:6073-6105.

[40] Leinster, T. (2004) *Higher Operands, Higher Categories* (Cambridge: Cambridge University).

[41] Crans, S. (2000) On braidings, syllepses and symmetries. *Cahiers Topologie Geom. Differentielle Categ*. **41**(1):2-74.

[42] Klein, D. J., Ivanciuc, T., Ryzhov, A. and Ivanciuc, O. (2008) Combinatorics of reaction-network posets. *Comb. Chem. High Throughput Screen*. **11**:723-733.

[43] Ivanciuc, T., Ivanciuc, O. and Klein, D. J. (2011) Network-QSAR with reaction poset quantitative superstructure–activity relationships (QSSAR) for PCB chromatographic properties. *Curr. Bioinform*. **6**:25-34.

[44] Restrepo, G., Brüggemann, R. and Klein, D. J. (2011) Partially ordered sets: Ranking and prediction of substances' properties. *Curr. Comput.-Aided Drug Des*. **7**:133-145.

[45] Smoluchowski, M. (1916) Drei vorträge über diffusion, Brownsche molekularbewegung und koagulation von kolloidteilchen. 1. *Phys. Z*. **17**:557-571.

[46] Smoluchowski, M. (1916) Drei vorträge über diffusion, Brownsche molekularbewegung und koagulation von kolloidteilchen. 2. *Phys. Z*. **17**:585-599.

[47] Debye, P. (1942) Reaction rates in ionic solutions. *Trans. Electrochem. Soc*. **82**:265-272.

[48] Berg, H.C. and Purcell, E. M. (1977) Physics of chemoreception. *Biophys. J*. **20**:193-219.

[49] KIRBY, A. J. (1993) Molecular recognition of transition states. *Philos. Trans. R. Soc. London, Ser. A* **345**:67-76.

[50] WILLIAMS, I. H. (2010) Catalysis: Transition-state molecular recognition? *Beilstein J. Org. Chem.* **6**:1026-1034.

[51] WHITNEY, C. K. (2012) On the several molecules and nanostructures of water. *Int. J. Mol. Sci.* **13**:1066-1094.

[52] SZEKELY, J., EVANS, J. W. and SOHN, H. Y. (1976) *Gas–Solid Reactions* (New York, NY: Academic), pp. 125-161.

[53] GARZA-GARZA, O. and DUDUKOVI≈, M. P. (1982) A variable size grain model for gas–solid reactions with structural changes. *Chem. Eng. J.* **24**:35-45.

[54] GEORGAKIS, C., ARIS, R. and AMUNDSON, R. (1977) Studies in the control of tubular reactors-I. General considerations. *Chem. Eng. Sci.* **32**:1359-1369.

[55] PATISSON, F., FRANÇOIS, M. G. and ABLITZER, D. (1998) A non-isothermal, non-equimolar transient kinetic model for gas–solid reactions. Chem. Eng. Sci. **53**:697-708.

[56] NIKSIAR, A. and RAHIMI, A. (2009) A study on deviation of noncatalytic gas–solid reaction models due to heat effects and changing of solid structure. Powder Technol. **193**:101-109.

[57] KHOSHANDAM, B., KUMAR, R. V. and JAMSHIDI, E. (2005) Simulation of non-catalytic gas–solid reactions: Application of grain model for the reduction of cobalt oxide with methane. *Min. Proc. Extr. Met.* **114**:10-22.

[58] GARIDEL, P. (2012) (2012) Right angle light scattering protein thermostability screening: From research to development. *Spectrosc. Eur.* **24**(1):13-18.

[59] HLAVACEK, W. S., FAEDER, J. R., BLINOV, M. L., POSNER, R. G., HUCKA, M. and FONTANA, W. (2006) Rules for modeling signal-transduction systems. *Sci. STKE* **344**:re6–1-18

[60] DOPAZO, J. (personal communication).

[61] SCHAD, E., TOMPA, P. and HEGYI, H. (2011) The relationships between proteome size, structural disorder and organic complexity. *Genome Biol.* **12**:R120-1-13.

[62] MITCHELL, K. J. (2011) The genetics of neurodevelopmental disease. *Curr. Opin. Neurobiol.* **21**:197-203.

[63] MITCHELL, K. J. and PORTEUS, D. J. (2011) Rethinking the genetic architecture of schizophrenia. *Psychol. Med.* **41**:19-32.

[64] MITCHELL, K. J. (2012) What is complex about complex disorders? *Genome Biol.* **13**:237-1-11.

[65] WANG, X., WEI, X., THIJSSEN, B., DAS, J., LIPKIN, S. M. and YU, H. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**:159-164.