Journal of Mathematical Chemistry 9(1992)351–357

IMPACT OF REGULARITY IN STRUCTURE–PROPERTY RELATIONSHIPS

Christiane MERCIER, Yves SOBEL and Jacques-Emile DUBOIS
ITODYS de l'Université Paris 7, accocié au C.N.R.S., 1 rue Guy de la Brosse, 75005 Paris, France

Received 25 November 1991

Abstract

The regular evolution of properties with structural modification is quantitatively formulated. It is defined in a structural space which is exhaustive, ordered, flexible and explicit. It is detected along the ordered pathways of structural filiations by inference tools, which take into account the experimental precision and proceed by a heuristic modulation of the initial representation space. The introduction of nuances in the quantitative formulation of regularity leads to diverse tools for setting up and exploiting the relationships. These latter constitute a prediction law, which is both simple and as precise as the measurements. The detected regularities make for a generalization of certain structural effects and suggest a fruitful interpretation. Zeroth- and first-order regularities, characteristic of strongly linear variation, are used to safely extrapolate the prediction range and to orient experimental planning.

1. Introduction

The purpose of structure–property relationship studies is to identify the effects of structural changes on a given property and to express them by a mathematical model. The notion that the information evolves regularly with structural modifications underlies any search for such a relationship. Only when the data are intrinsically regular is it possible to obtain a simple prediction law whose validity can be inferred beyond the range explored.

The detection of this regularity, and even its quantitative formulation, is difficult in a structural space since the variables are not metric. They do not even belong to a nominal category, where they would be either exclusive or independent. Moreover, the simplified representation of this structural space by a derived metric one (topological indices [1,2], physicochemical parameters [3]) a priori deletes some structural aspects, whose influence cannot then be investigated [4,5]. We thus looked for a representation resolving these difficulties by being: (a) *exhaustive* to make it possible to detect all the structural modifications influencing the behaviour and *ordered* so as to define a reference for the regularity, (b) *flexible* to make it adaptable to any structural and experimental field, and (c) *explicit* so as to identify the structural features responsible for the regularities.

In the DARC/PELCO method [6,7], whose structural variable answers the above criteria [5], the search for the best structure–property relationship is based on the detection of intrinsic regularities of phenomena evolving along the ordered pathways of structural filiations [8].

2. Regularity concept and detection tools

The DARC/PELCO variable [5] uses an ordered topology for describing the structures and the population which contains them. This dual representation leads to an exhaustive description of all the sites (atoms, bonds, stereochemical data) of a structure S, as well as its precise location in an organized population of affiliated structures, called a hyperstructure HS. In this space, the filiations between structures are organized in ordered pathways along which certain structural effects might vary in a regular manner. By evaluating the variations of a property along these structural filiations, it is possible to determine whether these effects are regular or not.

Regularity is defined by analogy with the trend towards linearity in a multidimensional vectorial space [8]. The variation of a property along a hyperstructure path has zeroth-order regularity when it tends to a plateau or to a horizontal asymptote, and first-order when it tends to a slope or an oblique asymptote (fig. 1). The regularity areas R_0 and R_1 are characterized by groups of sites whose influence on property is the same.

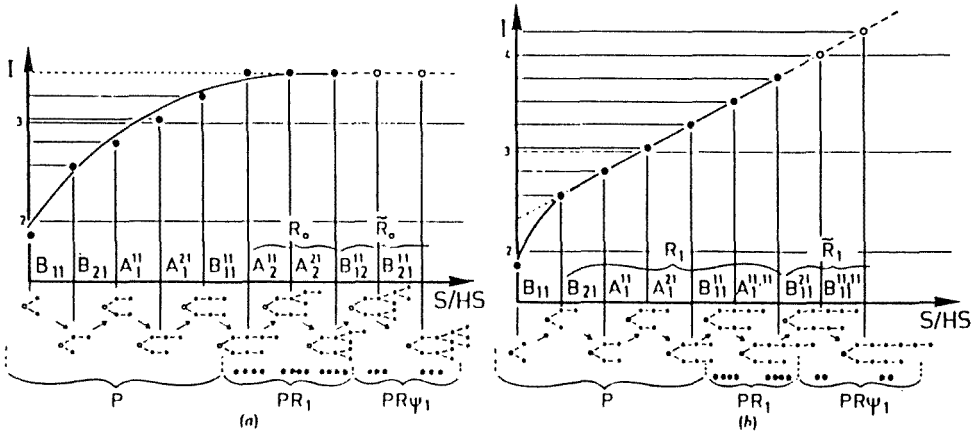


Fig. 1. Zeroth- and first-order regularities. Anesthesic activity evolving along a hyperstructure path reaches a horizontal asymptote (a) when the filiations correspond to branching sites, and an oblique asymptote (b) when the filiations correspond to chain-lengthening sites [5].

Several inference tools are used to detect these regularities. They consider in turn homogeneous series of the same type $(s_1, \ldots, s_i, \ldots)$ according to linear or

concentric homology [9,10], chromatism variation reflecting electronic homology [11,12] or a physicochemical scale (π, σ, E_s). They examine the value of the information perturbation $p(s_i)$ associated with the adjunction of each site s_i along a homogeneous structural filiation.

Δ being the experimental precision:

if $|p(s_i)| \leq \Delta$, then s_i belongs to R_0;

otherwise, if $|p(s_i)| \gg \Delta$ and

$|p(s_{i+1}) - p(s_i)| \leq \Delta$, then s_i and s_{i+1} belong to R_1.

The sites belonging to regularity areas are grouped in equivalence sites which are used to optimize the structure–property relationship.

3. Optimal structure–property relationship search

The DARC/PELCO search for the best model [5] is based on the detection of regularities in the multidimensional space *S/HS/I* representing the data (fig. 2). In this space, the experimental value related to a compound is equal to the sum of the reference value and of the information perturbations $p(s_i)$ associated with the site adjunctions leading to its structure. Regularity is first sought on an intermediate level [5], where it is assumed that the information perturbations are additive. This is based on the notion of *average perturbation* $\bar{p}(s_i)$. Finding their values amounts to seeking a hyperplane of *S/HS/I* space as near as possible to all the experimental points. When the perturbations are not additive, the hyperplane presents *singularities* which have to be taken into account. *Stronger regularities* are detected by applying the inference tools and lead to a model impervious to fluctuations arising from experimental uncertainty. The initial representation space, which contains all the sites used to describe the structures, is progressively modulated by introducing complex sites up to the optimal representation. Interaction sites account for some additivity deviations, while equivalence sites reflect regularities. At each step, the new variables to be introduced are deduced from the previous treatment and the new relationship is set up by multiple regression analysis. The most significant parameters are selected by a stepwise regression procedure.

The regularity hypothesis is often necessary in order to avoid the model to account for fluctuations arising from experimental errors. When experimental data exhibit regularity, this methodology is able to reflect it optimally and thus leads to a simpler and more general relationship. These qualities facilitate its interpretation and extend its prediction range.

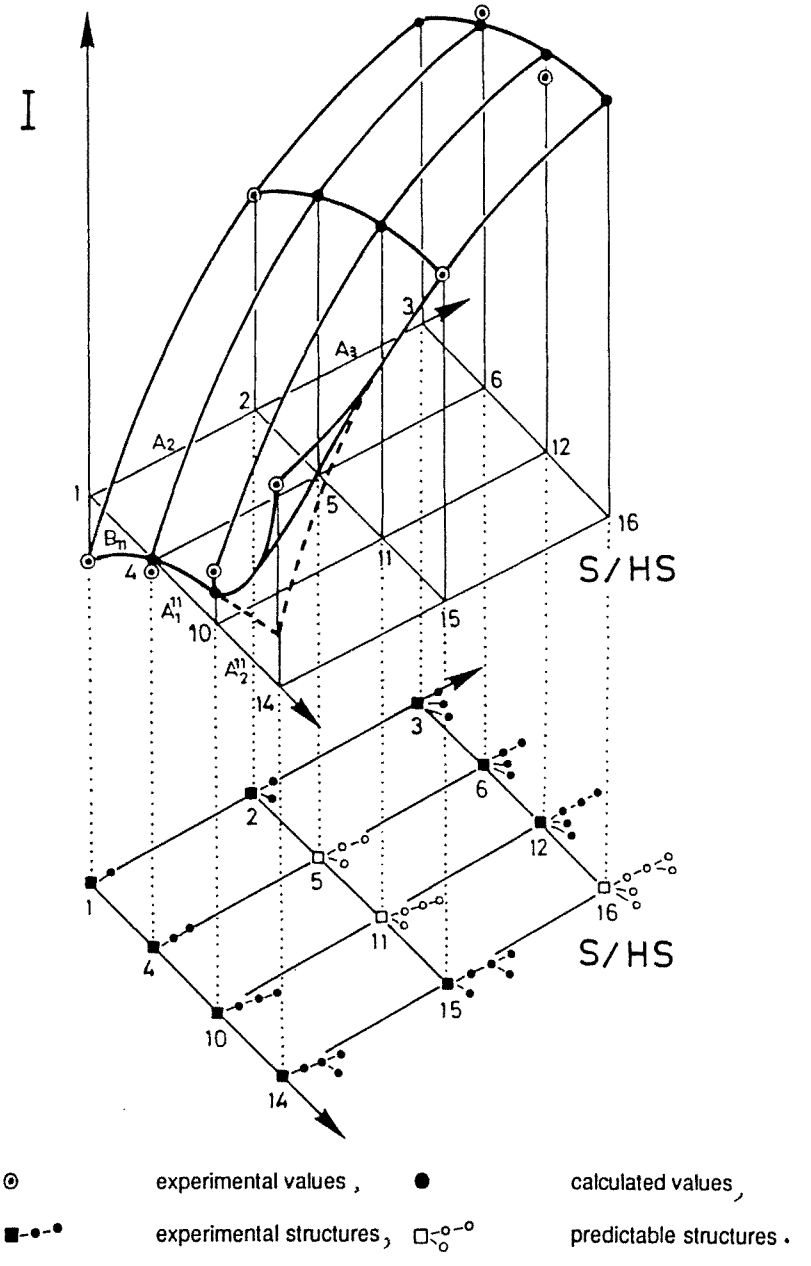


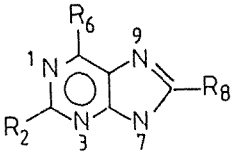
Fig. 2. Hyperplane of the optimal structure–property relationship. The set of the calculated values of the experimental and predictable structures furnishes a visual display of how the property develops along the hyperstructure filiations for glucuronide ability of alcohols. This makes it possible to define the areas of regular activity variation with structural modifications.

4. Interpretation and safe extension of the prediction range

4.1. INTERPRETATION

The regular influence of certain structural modifications, grouped in equivalence sites, makes for a *generalization of certain effects*. Thus, the partition coefficient of alcohols, ethers and amines is regularly enhanced by chain-lengthening (+0.54) and branching (+0.38). It is lowered by cyclization (−0.33). This structural influence relies on a model [10] based on a key population of measurements isolated from 117 values of partition coefficients for 60 compounds. The Structure–Log *P* relationship ($R^2 = 0.99$, $F = 2184$, $s = 0.09$) explains the property with four parameters, the three mentioned above and an additional one reflecting the shift in behaviour between alcohols and ethers.

Analysing the pK_a of purines [12], we set up a five-variable model for 40 anionic pK_as ($R^2 = 0.99$, $F = 4748$, $s = 0.05$) and a four-variable model for 34 cationic pK_as ($R^2 = 0.99$, $F = 3966$, $s = 0.07$). The 2- and 6-substituents have a similar influence, especially for cations, and the 8-substituents lower the pK_a of anions to a greater extent. This is consistent with the fact that N_1 is the site of protonation in purine and that the negative charge is shared between N_7 and N_9.



It is important to have good predictive models for the basic physicochemical properties, such as log *P* or pK_a, especially because the structural influence sometimes suggests an interpretation in terms of these effects. Anesthesic activity was found to be regularly enhanced by chain-lengthening and branching up to a saturation threshold [8]. This influence can be explained by a second degree function of their partition coefficients. Indeed, when we applied a topo-hydrophobic model to this series, we found that the partition coefficient was the only significant parameter.

Exceptional contribution values can suggest new fields of investigation towards an area of higher activity. When we investigated the PNMT inhibitory potency of 109 phenylalkylamines, four really significant variables were detected [11]. The sites of direct substitution on the aromatic nucleus constitute the major source of increasing activity. Their influence grows regularly with the nature of chromatism on each ortho, meta and para position, following the sequence CH_3, F, Cl, Br, I, CF_3 and suggesting an electronic interpretation. Exceptional modifications, grouping a branching site, a cyclization and a double bond, have the same activating influence. They can begin families and point out research directions that would be worthwhile exploring.

Regularity areas are delimited by *frontiers* [5]:

- zero and equivalence frontiers, F_0 and F_1, beyond which site influence on property is zero or equivalent;
- inhibiting frontiers $F_i(s)$, beyond which sites have no effect when the site s is occupied.

These frontiers make for a general interpretation. The five frontiers detected in the alcohol metabolism model [9] ($R^2 = 0.98$, $F = 379$, $s = 0.09$) reflect the fact that the fifth carbon introduced into the environment for secondary and tertiary alcohols, and the sixth carbon for primary alcohols have no effect.

4.2. SAFE EXTENSION OF THE PREDICTION RANGE

The reliable prediction area of the relationship, called "Proference", consists of structures whose predicted property is obtained by interpolation. Proference includes all the structures generated from the studied population and which do not belong to it [5,13]. Regularity enhances the reliability of the predictions at a given level of structural interpolation, this latter reflecting the surrounding of the predictable structures by experimental ones. Regularity is best used to safely extend the prediction range beyond the field explored. This range, called "Pseudo-proference", consists of structures whose predicted property is based on extrapolations of the regularities detected in the experimental population. It is constructed by prolongation of the regularity areas, reflected by an extension of the population trace. It includes the structures generated on the enlarged trace and which do not belong to the proference [5,8]. The reliability of the new predictions depends on the quality of the regularity hypothesis (fig. 3).

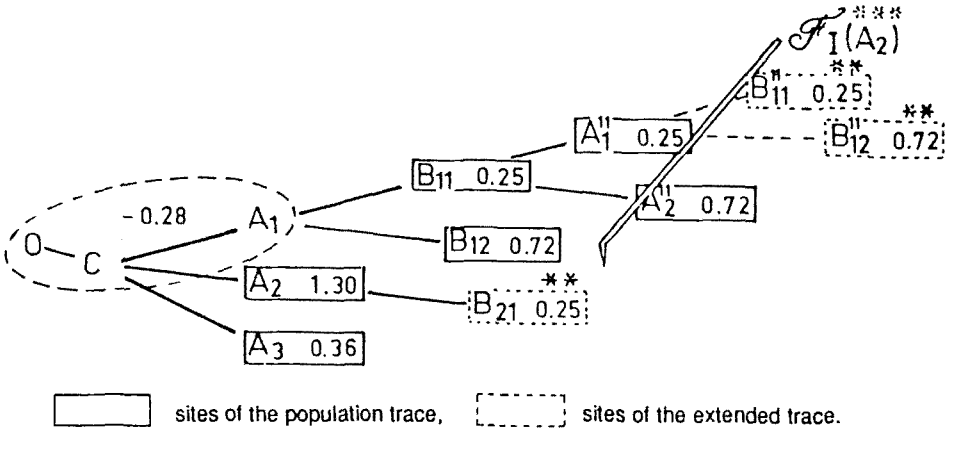


Fig. 3. Extending regularity areas and predicting by extrapolation. Regularities are used to assign a perturbation value to similar sites that depart from the population trace. Predictions deduced from first-order regularity have ** reliability, those deduced from zeroth-order regularity have *** reliability.

The extrapolated predictions are used as simulation tools for structural elucidation or for constructing intermediate structural variables used in LFER and QSAR (pK, log P, σ, E_s). They are very useful to plan experiments for elucidating mechanisms or to search for compounds of maximal activity and minimal toxicity. In the anticholinergic series, the extension of the active structural range and experiments on relevant compounds have made it possible to determine the substructure responsible for the activity. It has led to the discovery of particularly effective molecules exhibiting an activity ten times higher than the best of the previous compounds [14].

References

[1] L.B. Kier and L.H. Hall, Molecular connectivity in chemistry on drug research, in: *Medicinal Chemistry*, Vol. 14, ed. G. de Stevens (Academic Press, London, 1976).
[2] O. Mekenyan and D. Bonchev, Acta Pharma. Jug. 36(1986)225.
[3] C. Hansch, Quantitative structure–activity relationships in drug design, in: *Drug Design*, Vol. 1, ed. E.J. Ariens (Academic Press, London, 1971).
[4] C. Mercier and J.E. Dubois, Eur. J. Med. Chem. 14(1979)415.
[5] C. Mercier, Y. Sobel and J.E. Dubois, Ann. Math. Chem., in press.
[6] J.E. Dubois, DARC system in chemistry, in: *Computer Representation and Manipulation of Chemical Information*, ed. W.T. Wipke, S. Heller, R. Fellmann and E. Hyde (Wiley, New York, 1974), p. 239.
[7] J.E. Dubois, D. Laurent and A. Aranda, Perturbation of environments which are limited, concentric and ordered, J. Chim. Phys. 11–12(1973)1608; 1616.
[8] J.E. Dubois, Y. Sobel and C. Mercier, C.R. Acad.. Sci. Paris, série II, 292(1981)783.
[9] C. Mercier, V. Fabart, Y. Sobel and J.E. Dubois, J. Med. Chem. 34(1991)934.
[10] C. Mercier, D. Ouaknin and J.E. Dubois, to be published.
[11] C. Mercier, Y. Sobel and J.E. Dubois, Eur. J. Med. Chem. 16(1981)473.
[12] C. Mercier, O. Mekenyan, J.E. Dubois and D. Bonchev, Eur. J. Med. Chem. 26(1991)575.
[13] J.E. Dubois, C. Mercier and Y. Sobel, C.R. Acad. Sci. Paris, 289C(1979)89.
[14] C. Mercier, G. Trouiller and J.E. Dubois, Quant. Struct.–Act. Rel. 9(1990)88.