Aplicació de metodologies quimiomètriques a l'estudi de l'efecte del solvent sobre els aspectes termodinàmics i estructurals dels equilibris àcid-base dels polinucleòtids

Anna de Juan Capdevila

DEPARTAMENT DE QUÍMICA ANALÍTICA DE LA UNIVERSITAT DE BARCELONA

PROGRAMA DE DOCTORAT: QUÍMICA ANALÍTICA DEL MEDI AMBIENT I DE LA POL·LUCIÓ. (BIENNI 1988-1990)

APLICACIÓ DE METODOLOGIES QUIMIOMÈTRIQUES A L'ESTUDI DE L'EFECTE DEL SOLVENT SOBRE ELS ASPECTES TERMODINÀMICS I ESTRUCTURALS DELS EQUILIBRIS ÀCID-BASE DELS POLINUCLEÒTIDS

Memòria presentada per Anna de Juan i Capdevila per optar al grau de Doctor en Ciències Químiques.

Directors: Enric Casassas i Simó i Gemma Fonrodona Baldajos.

Barcelona, juliol de 1997.

Enric Casassas i Simó, catedràtic emèrit del Departament de Química Analítica de la Universitat de Barcelona, i Gemma Fonrodona Baldajos, professora titular del mateix departament,

CERTIFIQUEN:

que la present memòria, que duu per títol: "Aplicació de metodologies quimiomètriques a l'estudi de l'efecte del solvent sobre els aspectes termodinàmics i estructurals dels equilibris àcid-base dels polinucleòtids", ha estat realitzada sota la nostra direcció per Anna de Juan i Capdevila al Departament de Química Analítica de la Universitat de Barcelona i que tots els resultats presentats són fruit de les experiències realitzades per l'esmentada doctoranda.

Barcelona, juliol de 1997.

Enric Casassas i Simó

Gemma Fonrodona Baldajos

La capacitat d'agraïment és probablement un dels trets més diferencials i positius de la natura humana. Reuneix alhora la nostra part més racional, que valora l'ajut rebut en un moment determinat, i la nostra part més instintiva, que sent el suport emocional d'aquells que ens donen la mà quan més ens cal.

La necessitat d'agrair després d'un llarg camí és vivíssima, gairebé essencial, però la memòria és flaca i la por a oblidar, ni que sigui un sol moment o una sola persona, tenalla la mà i l'esperit de qui vol expressar aquest sentiment. És per això que no vull, no puc incloure noms. Tots vosaltres, els qui m'heu fet créixer i m'heu fet costat personalment i científicament durant tots aquests anys, esteu aquí, en aquest full i en el meu pensament, i és a vosaltres a qui demano que vulgueu fer vostre un meu missatge tan curt com sincer:

-Gràcies per ésser-hi.

Al Dr. Enric Casassas

INDEX

II. RESULTS.

CHAPTER 4. ON THE MICROSCOPIC DESCRIPTION OF SOLVENT SPACE

CHAPTER 5. GOING FROM PURE SOLVENTS TO SOLVENT MIXTURES: THE WATER-DIOXANE EXAMPLE.

5.1. Microscopic characterization of the mixture and determination of acid-base equilibria.

5.2. Modelling of solvent-dependent processes in water-dioxane mixtures: proposals for the establishment of Linear Free Energy Relationships (LFER).

OBJECTIVES

The main goal of this project is the qualitative and quantitative description of all the thermodynamical and conformational transitions related to the acid-base behaviour of several polynucleotides in biological environments of low polarity. The emulation of these special environments has been carried out by using water-dioxane mixtures that keep the aqueous nature of the biological media and present the desired low polarity due to the features of their cosolvent.

Owing to the complexity associated with the macromolecular nature of the polynucleotides and with the mixed character of the solvent used, some fundamental research must be carried out before facing specifically the research about the acid-base polynucleotide behaviour in water-dioxane mixtures. These previous studies include, on one hand, the detailed characterization of the water-dioxane mixtures and the interpretation of their effect on the acid-base behaviour of single solutes and, on the other hand, the study of the monomeric units of the polynucleotides in these mixtures.

In all the work performed, the careful treatment of the experimental data has been a constant concern. Special attention has been focused on the establishment of Linear Solvation Energy Relationships (LSER), behaviour models that relate the solute behaviour to the solvent effect, and on the interpretation of the multivariate data coming from the monitoring of the macromolecular equilibria of polynucleotides. The former problem has been tackled by using different kinds of hard-modelling and soft-modelling methods, whereas the latter has been solved with the application of curve resolution methods, which do not need the postulation of any chemical model to interpret the variation of the different species in solution.

STRUCTURE OF THE WORK PRESENTED

The series of articles included in this project shows the whole process followed in the study of the acid-base behaviour of polynucleotides in low polarity environments, which begins with the selection and assessment of the solvent descriptors used in the later characterization of the water-dioxane mixtures (article I), follows with the microscopic characterization of these hydroorganic solvents and with the interpretation of their effect on the acid-base behaviour of simple solutes (articles II-VI) and finishes with the multivariate monitoring and the interpretation of the acid-base behaviour of the polynucleotides in these solvent mixtures (articles VII-XI).

Thus, article I presents an overview of the evolution in the solvent description, enhancing the transition from the macroscopic properties (suitable to characterize the bulk solvent) to the microscopic parameters (more appropriate to characterize the solvent features in the cybotactic region around the solutes). The microscopic descriptors proposed by Kamlet et al. to quantify the hydrogen-bonding and the polarity solute/solvent interactions are used to structure the global solvent space in solvent groups built according to the similarity of the microscopic features of their members. This solvent classification is later compared with the classical scheme of Snyder.

The gaps among the groups of pure solvents are filled by the infinite number of possible solvent mixtures. The water-dioxane system covers a wide zone of the solvent space and has been selected for the later studies in this project. Articles II, III and IV include the experimental determination of the microscopic properties related to several hydroorganic mixtures formed by changing the proportion of dioxane from 0 to 100 %

(v/v) and the determination of some protonation constants of simple solutes dissolved in the mixtures previously characterized. Simple correlations between different sets of microscopic parameters, between microscopic parameters and solvent composition and between protonation constants and solvent descriptors are presented and interpreted. Articles V and VI are devoted to the establishment of more accurate behaviour models to explain the solvent effect on the solute property under study. Two possible alternatives are proposed to build these LSER models: the first involves a robust strategy to fit the experimental data to a postulated chemical model (hard-modelling approach) and the latter combines the use of Factor Analysis (FA) and Target Factor Analysis (TFA), which do not need the use of an initial chemical expression and transform an abstract model (FA) into a chemically meaningful LSER (TFA) (soft-modelling approach).

Once the water-dioxane system has been characterized and some simple acid-base processes have been analyzed in it, the study of macromolecular processes, such as the acid-base behaviour of polynucleotides, can be carried out. Handling the experimental output coming from these experiments is not a trivial task; therefore, some research has been previously focused on the improvement of the currently used curve resolution method, the Alternating Least Squares method (ALS), through the proposal and assessment of new constraints to be used in the iterative resolution procedure (article VII). The performance of this method has been compared with the Trilinear Decomposition (TLD), another curve resolution method with a different background (article VIII). Simulated data with a large variety of features and real data have been used in both chemometric studies. After having confirmed the suitability of the improved ALS method to handle data with the features of the multivariate output coming from the monitoring of macromolecular equilibria, articles IX-XI explain in detail the thermodynamical and conformational transitions related to the acid-base equilibria of the polyuridylic acid (polyU), polycitydylic acid (polyC) and polyadenylic acid (polyA) in water-dioxane mixtures and compares these results with others coming from previous studies carried out in aqueous solution. Articles IX and X are specifically oriented to the description of the pH-dependent transitions of the polyU-H and the polyC-H systems, respectively, and of their related cyclic monomeric nucleotides. Article XI shows a compilation of the results shown in articles IX and X and makes an intercomparison between the chemical conclusions related to these two polynucleotides and to the polyA-H system. This last article stresses as well the big potential of the combined use of experimental multivariate monitoring and curve resolution techniques for the study of biomacromolecular equilibria through the systematic exposition of all the different kinds of information that can be obtained when this approach is followed.

ARTICLES PRESENTED

I. Solvent classification based on solvatochromic parameters. A comparison with the Snyder approach.

A. de Juan, G. Fonrodona and E. Casassas.

Trends in Analytical Chemistry, 16 (1997) 52-62.

II. Correlation of acid-base properties of solutes with the polarity parameters and other solvatochromic parameters of dioxane-water mixtures.

E. Casassas, G. Fonrodona and A. de Juan.

Inorganica Chimica Acta, 187 (1991) 187-195.

III. Determinación del parámetro de polaridad-polarizabilidad π^* y correlación de éste con $E_T(30)$ para mezclas dioxano-agua.

E. Casassas, G. Fonrodona and A. de Juan.

Anales de Química, 87 (1991) 611-615.

IV.Solvatochromic parameters for binary mixtures and a correlation with equilibrium constants. Part I. Dioxane-water mixtures.

E. Casassas, G. Fonrodona and A. de Juan.

Journal of Solution Chemistry, 21 (1992) 147-162.

V. Assessment of solvent parameters and their correlation with protonation constants in dioxane-water mixtures using factor analysis.

E. Casassas, G. Fonrodona, A. de Juan and R. Tauler.

Chemometrics and Intelligent Laboratory Systems, 12 (1991) 29-38.

VI. Factor Analysis applied to the study of the effect of solvent composition and of the inert electrolyte nature on the protonation constants in dioxane-water mixtures.

E. Casassas, N. Domínguez, G. Fonrodona and A. de Juan.

Analytica Chimica Acta, 283 (1993) 548-558.

VII. Assessment of new constraints applied to the Alternating Least Squares (ALS) method.

A. de Juan, Y. Vander Heyden, R. Tauler and D.L. Massart.

Analytica Chimica Acta, (1997) (in press).

VIII.Comparison between the Trilinear Decomposition (TLD) and the Alternating Least Squares (ALS) methods for the resolution of three-way data sets.

A. de Juan, S.C. Rutan, R. Tauler and D.L. Massart.

Submitted to *Chemometrics and Intelligent Laboratory Systems*.

IX.Application of a self-modeling curve resolution approach to the study of solvent effects on the acid-base and copper(II)-complexing behaviour of polyuridylic acid.

A. de Juan, G. Fonrodona, R. Gargallo, A. Izquierdo-Ridorsa, R. Tauler and E. Casassas.

Journal of Inorganic Biochemistry, 63 (1996) 155-173.

X. Three-way curve resolution applied to the study of solvent effect on the thermodynamic and conformational transitions related to the protonation of polycytidylic acid.

A. de Juan, A. Izquierdo-Ridorsa, R. Gargallo, R. Tauler, G. Fonrodona and E. Casassas.

Analytical Biochemistry (1997) (in press).

XI.A soft-modeling approach to interpret thermodynamic and conformational transitions of polynucleotides.

A. de Juan, A. Izquierdo-Ridorsa, R. Tauler, G. Fonrodona and E. Casassas.

Biophysical Journal (1997) (accepted for publication).

I. GENERAL INTRODUCTION.

CHAPTER I.

SOLVENTS AND SOLUTE/SOLVENT INTERACTIONS.

1.1 A historical review of the description of solvents: from bulk properties to microscopic parameters.

Progress in all fields of chemistry has been closely connected to the developments in the understanding of solvents and solutions. From earliest times, there have always been researchers for whom the solvent was much more than "container of solutes". For them, the solvent was seen as an active element responsible for the evolution of the chemical processes occurring in solution (Boerhaave,1733), (Reichardt,1988).

One of the main concerns of the alchemists was the identification of a universal solvent, the so-called alkahest. For some centuries, numerous attempts were made at revealing this utopian substance. Though this specific objective was never attained, the discovery of new solvents and processes helped to increase the general understanding of the role of solvents and solutes in solution. Certain elementary rules, such as "like solves like" appeared during this period; however, the concept of solution remained vague and the loss of the nature of a substance with dissolution was widely accepted. In the 17th century, Van Helmont was the first to question this theory. He argued that the substance as such remained in solution, though in an "aqueous" form, and believed it could be recovered from solution after applying the suitable procedure. This alternative gradually acquired general acceptance by the end of the 19th century, supported by Van't Hoff's theory of osmotic pressure and by Arrhenius's theory of electrolyte dissociation.

The first solvent effects reported also date from the end of the 19th century. Thus, Berthelot and Péan de Saint Gilles were the first to notice the solvent effect in the rate of chemical reactions in 1862 (Berthelot,1862). Within the kinetic field, Menshutkin's contribution in 1890 concerning the solvent effect on the alkylation of tertiary amines with haloalkanes became a point of reference and many fundamental statements in this study, such as "a chemical reaction cannot be separated from the medium in which it is performed" remain valid (Menshutkin,1890). The influence of solvents in chemical equilibria was revealed separately by Claisen (Claisen,1896), Knorr (Knorr,1896) and Wislicenus (Wislicenus,1896) in1896 from studies conducted on the keto-enol tautomerism of 1,3-dicarbonyl compounds.

The key to explaining the solvent effect in all chemical phenomena lies in a good description of the solvent itself, thus leading to a simpler interpretation of both the nature and extent of the possible interactions performed on the solutes. For a long time, the solvent was considered a nonstructured dielectric continuum and physical constants such as the refractive index (n) and the relative permittivity or dielectric constant (ε) were supposed to be the best properties by which it could be described. These macroscopic properties, while highly suitable for characterizing the bulk solvent, do not take into account the specific solute/solvent interactions and often fail to correlate variations in solute properties caused by solvent effects. Indeed, the solvent around solutes can no longer be considered as a uniform medium, but as a structured discontinuum consisting of individual solvent molecules that interact specifically with each other and with solutes.

The complete description of a solvent must then include both macroscopic and microscopic parameters, the former mainly related to nonspecific solvent/solvent interactions (e.g., electrostatic contributions) and the latter to the quantitation of specific solute/solvent interactions (e.g., hydrogen-bonding). Whereas the experimental measurement of macroscopic properties is an easy task, there is no direct instrumental access to the microscopic environments. Therefore, microscopic descriptors must be obtained using alternative experimental strategies to those applied in the determination of macroscopic physical properties.

Given the close relationship between solutes and solvent, experimental measurements of solute properties provide information about the solute itself and about the surrounding solvent. The possibility of obtaining indirect information about the solvent around the solutes by measuring the effect that the former has on certain well-known and strongly solvent-dependent processes is the common basis for many empirical scales of solvent microscopic parameters (Reichardt,1988). In these processes, the solute involved behaves as a probe in the solvation shell reflecting changes in the surrounding solvent through variations in its absorption spectra or in some thermodynamic or kinetic parameters.

The first empirical scale was defined by Winstein in 1948 (Winstein,1948). He proposed a Y parameter to represent the "solvent ionizing power", which was built from the measure of the kinetic constant related to the S_N1 solvolysis of 2-chloro-2-methylpropane. Additional examples of scales with kinetic reference processes are the X scale (Gielen,1963) or the four parameter approach proposed by Swain et al. (Swain,1955). Equilibrium processes have also been employed in the establishment of empirical parameters (Drago,1965), (Maria,1985). Gutmann's donor number, DN, is a well known example proposed as a measure of the solvent Lewis basicity (Gutmann,1966); DN is defined as the negative value of the molar enthalpy for the adduct formation between antimony pentachloride and electron-pair donor solvents measured in highly diluted solutions of 1,2-dichloroethane as inert solvent.

Despite the proven value of empirical parameters based on kinetic or thermodynamic properties, empirical scales using spectroscopic data are more generally employed because of the overall simplicity in the recording of the related experimental measurement. Indeed, solutes absorption spectra show changes in the position, intensity and shape of their absorption bands due to the solvent-dependent alteration of the energy difference between the ground and the excited states of their chromophores. In 1951, Brooker suggested the potential usefulness of solvatochromic dyes in quantifying the solvent microscopic properties (Brooker,1951). Seven years later, Kosower set up the first empirical scale based on spectroscopic properties (Kosower, 1958a; 1958b; 1958c). His Z polarity scale took the charge transfer transition of 1-ethyl-4-(methoxycarbonyl)pyridinium iodide as spectroscopic reference process.

The many solvatochromic scales which exist are built from UV, visible or near-IR spectroscopic measurements and often take solvent-sensitive $n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$ or charge transfer transitions as reference spectroscopic processes. Though many compounds exhibit spectral changes due to solvent effects, not all of them are suitable for use as reference solutes of an empirical solvent scale. An ideal solvatochromic probe should fulfil the following requirements:

- **High sensitivity to changes in the surrounding medium.** The solvent-dependent bathochromic or hypsochromic shifts in the position of the wavelength corresponding to the absorption band maximum must be as large as possible and always significantly larger than the instrumental accuracy associated with the wavelength location.
- **Large molar absorptivities.** Solvatochromic indicators are used to describe solute/solvent interactions. Therefore, the molar absorptivity of these substances must be large enough in all the solvents studied to allow one to work with highly diluted solutions, in which the existence of spectral shifts induced by solute/solute interactions caused by the aggregation of probe molecules can be discarded.
- **Absorption band out of the solvent spectral region.** The estimation of the solvatochromic shifts is more accurate if the overlap between the absorption bands of the probe and the solvent is absent or as small as possible.
- **Solubility.** The solvatochromic probe must be soluble in solvents with many different features in order to obtain representative empirical scales.
- **Low reactivity.** The chemical nature of the solvent-sensitive chromophore must not be altered by any chemical reaction between the indicator and the solvent.

- **Stability.** The reference solute must be a crystalline substance with a perfectly defined chemical structure, stable in storage and in solution.
- **Availability.** The probe must be commercially available or, at least, easily synthesized.

The evident shortcoming of all the solvatochromic scales is the inherent assumption that the solvent behaviour around the probe molecule is the same as that around any other kind of solute. This oversimplification is, however, less dangerous than assuming that the bulk solvent properties describe the microscopic environment around the solutes. Some authors recommend the establishment of empirical scales using averaged values coming from several probes in order to minimize the influence of the different nature of each reference solute on the microscopic parameter. Nevertheless, this average will only be meaningful if the dispersion among the results provided by all the indicators considered is sufficiently small. If this is not the case, the use of single probe-based scales, whose solute/solvent interactions are clearly defined, will yield more valuable and sound conclusions about the evolution of solvent-dependent processes than a carelessly averaged parameter. As an additional recommendation, the use of empirical scales working with probes and reference processes which are as similar as possible to the solute and the solvent-dependent process under study will also reduce the differences between the microscopic environments around the probe and around the solute.

There are two marked approaches to the construction of solvatochromic empirical scales: the uniparametric and the multiparametric approaches. The former tries to include all the possible forms of solute/solvent interactions in a single parameter, whereas the latter associates each kind of solute/solvent interaction (i.e., hydrogen-bonding, polarizability,...) with a separate parameter and it is the combined use of all these specific parameters is which gives a global picture of the solvent. Both approaches can be equally useful owing to the complementary information they provide, which is related to the global intensity of the solute/solvent interactions on the one hand and to the nature of these interactions on the other.

Most of the uniparametric solvatochromic scales are generally defined as solvent polarity scales. In this context, the term polarity accounts for the overall solvation capability, which depends on all possible solute/solvent interactions which do not cause any definite chemical alteration in the reference solute (i.e., protonation, oxidation,...). For this reason, the probes used in these scales are usually substances with a large solvent-dependent solvatochromism, such as merocyanines or pyridinium N-phenolate betaine dyes, which are able to interact in many diverse forms with the solutes. Some examples of uniparametric approaches are the aforementioned Z scale, the $E_T(30)$ parameter proposed by Dimroth and Reichardt (Dimroth,1963), the RPM (Relative Polaritätsmass) scale of Dähne (Dähne,1975), the ϕ scale of Dubois (Dubois,1966) or the Py scale of Dong and Winnik (Dong,1982). From all the scales belonging to this family, $E_T(30)$ has been chosen as the uniparametric approach in this project due to the large negative solvatochromism of its reference probe, to the wide variety of solvents covered by this scale and to the good correlation found between $E_T(30)$ parameter and many other empirical parameters.

In contrast to the uniparametric scales, the solvatochromic indicators used in multiparametric approaches are selected giving priority to the specificity of the probe for a certain kind of solute/solvent interaction rather than to the magnitude of the exhibited solvatochromic shift. The several parameters forming a multiparametric approach are determined by taking one or more reference solutes sensitive to an only kind of solute/solvent interaction or, when this is not possible, by taking pairs of solutes whose difference in solvatochromic behaviour is due to only one source of solute/solvent interactions. Known examples of multiparametric approaches are those proposed by Koppel and Palm (Koppel,1971), Krigowski and Fawcett (Krigowski,1975) and Swain et al. (Swain,1983). The most successful and generally applied multiparametric approach is the so-called solvatochromic comparison method, proposed by Kamlet, Taft and Abboud. The different solute/solvent interactions are separated in the solvatochromic parameters α, β and π^* related to the hydrogen bond acidity, hydrogen bond basicity and polarity-polarizability, respectively (Taft,1976), (Kamlet,1976), (Kamlet,1977). Owing to the great specificity of each of the parameters above caused by the careful selection of the solvatochromic probes used in their determination and to their proven ability to describe variations of solvent dependent processes with linear models including weighted sums of the solute/solvent interactions represented by α, β and π*, the solvatochromic comparison method is the most widely used multiparametric approach and has been selected for use in this project.

Though the application of the most popular empirical scales dates back a number of decades, the microscopic solvent description is today an active research area. Current subjects in this field include a clearer understanding of the solvatochromic indicators being used (Dealencastro,1994), (Boggetti,1994), (Ramírez,1995), the proposal of new potential dyes with better solvatochromic properties or specifically oriented to the interpretation of some concrete solute processes (Scremin,1994), (Reichardt,1995a;1995b), (Effenberger,1995), (Albert,1996), (Lu,1996), the establishment of new empirical scales (Drago,1994), (Catalán,1995) or the adaptation of existing parameters so that they might describe microscopic environments in supercritical fluids (O'Neill,1993), (Sun,1995), (Schulte,1995), gas phase (Koppel,1994), solid phase (Park,1994), (Li,1995), (Spange,1996) or microheterogeneous solutions, such as micellar systems (Drummond,1986).

1.2. The $E_T(30)$ polarity scale: a uniparametric microscopic approach.

The $E_T(30)$ parameter was proposed by Dimroth and Reichardt in 1963 as a measure of solvent polarity (Dimroth,1963). The solvent-dependent process used in establishing this empirical scale is the $\pi \rightarrow \pi^*$ electronic transition of the solvatochromic dye 2,6-diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenolate and each $E_T(30)$ value is defined as the molar transition energy (in kcal/mol) of the reference dye dissolved in the solvent under study measured at 25° C and at normal pressure (Reichardt,1988). Thus, $E_T(30)$ is evaluated by introducing the wavenumber of the absorption band maximum related to the aforementioned transition in the following expression:

$$E_T(30)\ (\text{kcal/mol}) = N_A\ h\ c\ \nu_{max}$$

The dimensionless derivation of the $E_T(30)$ scale is the $E_T{}^N$ scale, where water and tetramethylsilane (TMS) are taken as extreme polar and non-polar reference solvents, respectively (Reichardt,1983).

$$E_T^N = \frac{E_T(\text{solvent}) - E_T(\text{TMS})}{E_T(\text{water}) - E_T(\text{TMS})}$$

The unusually large solvatochromic shift associated with the solvent-dependent transition employed in the establishment of the $E_T(30)$ scale ($\Delta\lambda$ = 357 nm when going from water to diphenyl ether) comes from the great difference in solvation between the ground and the excited state of the betaine dye. This difference stems from the polarity decrease associated with the charge transfer from the phenolate group to the pyridinium part of the molecule during the transition (see Figure 1.2.1.).

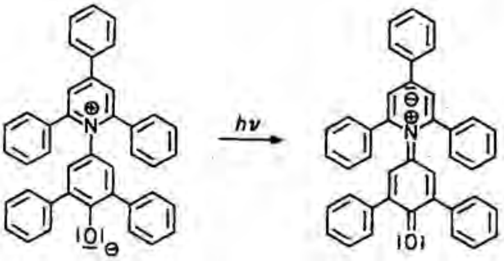


Figure 1.2.1. Chemical forms of the ground and excited states of Reichardt's betaine dye.

The chemical structure of Reichardt's betaine dye is responsible for the great sensitivity of this molecule to many different kinds of solute/solvent interactions. Thus, the large permanent dipole of the molecule allows the detection of dipole/dipole and dipole/induced dipole interactions, the large polarizable π-electron system, which includes 42 electrons, registers the dispersion interactions, and the phenolate group is a rather basic center, sensitive to the hydrogen-bonding acidity of the surrounding solvent. In contrast, the solvent hydrogen-bonding basicity cannot be detected because the positive charge of the pyridinium moiety is delocalized and sterically shielded, thereby preventing the establishment of interactions between the basic center of the solvent and the acidic center of the probe (Reichardt,1994).

Another outstanding quality of the $E_T(30)$ scale is the great compatibility of the reference dye with solvents with a wide range of features and behaviour. Direct measures of this parameter are reported for more than 350 pure solvents and many common solvent mixtures. In order to expand the $E_T(30)$ scale, Reichardt and col. have devoted considerable efforts to obtaining new indicators suitable for the determination of this parameter in solvents where the direct measurement of the $E_T(30)$ parameter is not possible. These substances are substituted betaines that keep the basic skeleton of the $E_T(30)$ reference dye and whose solvatochromic behaviour correlates perfectly with the behaviour of their parent molecule. The substituents introduced in these new dyes adapt the probe molecule to a concrete group of solvents; thus, a betaine with tert-butyl groups has been designed to deal with low polar solvents, in which the $E_T(30)$ dye is not soluble, and slightly acidic solvents can be analyzed by using a betaine with electron-withdrawing groups, which decrease the basic power of the phenolate group (Reichardt,1993).

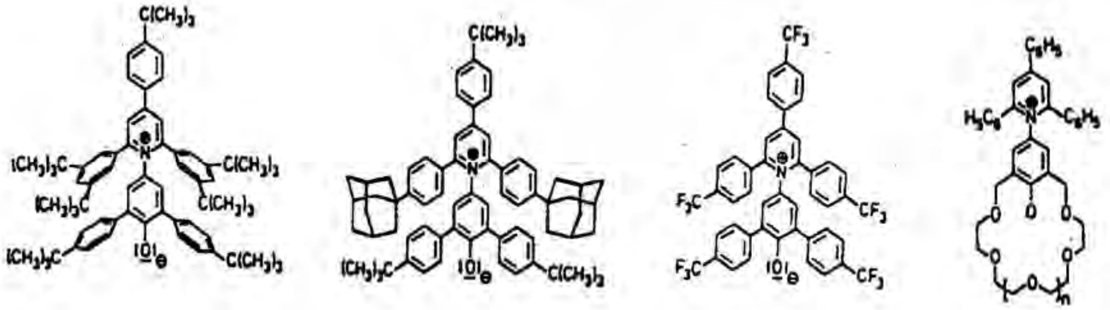


Figure 1.2.2. Some solvatochromic indicators obtained by modifying the basic $E_T(30)$ dye.

Despite the wide applicability of Reichardt's polarity scale, no $E_T(30)$ values can be experimentally determined, be it with the original $E_T(30)$ dye or with their derivatives, in very acidic solvents and in the gas phase. In the first case, the oxygen atom of the phenolate group protonates and the charge transfer solvatochromic absorption band disappears and, in the second case, the betaine dyes are not volatile enough to allow the $E_T(30)$ measurement in gas phase. The $E_T(30)$ values reported for very acidic solvents have then to be calculated from the mathematical expression correlating this empirical scale with the Z polarity scale, where experimental measurements for acidic solvents are available (Reichardt,1983). The gas phase $E_T(30)$ value, equal to 27.1 according to a recent review of Reichardt, is the average of the concordant results evaluated from theoretical and empirical relationships used in several independent approaches (Richert,1993), (Jano,1992), (Reichardt,1994).

1.3. The solvatochromic comparison method: a multiparametric microscopic approach.

The solvatochromic comparison method was born as an approach to unravel, quantify, correlate and interpret multiple interacting solvent effects on many kinds of chemical properties (Kamlet,1983). Kamlet, Abboud and Taft proposed a series of empirical microscopic solvent parameters, each of them specially designed to account for a specific solute/solvent interaction. Thus, α represents the solvent hydrogen-bond donor acidity (Taft,1976), β the solvent hydrogen-bond acceptor basicity (Kamlet,1976) and π* the solvent polarity/polarizability (Kamlet,1977). These parameters are usually determined from spectroscopic measurements, hence the generalized name of solvatochromic parameters.

The solvatochromic parameters were designed to be included in linear models used in explaining the solvent effect in the solute properties, the so-called Linear Solvation Energy Relationships (LSER). The most extended form of LSER proposed by Kamlet et al. has the form:

$$XYZ = (XYZ)_o + a\alpha + b\beta + s(\pi^* + d\delta) + h\delta_H^2 + e\xi$$

where XYZ is the solvent-dependent solute property under study, XYZ_o is the value of this property in a hypothetical inert solvent unable to interact with the solute, α, β and π* are the aforementioned solvatochromic parameters, δ is a polarizability correction term equal to 0 for nonchlorinated aliphatic solvents, 0.5 for chlorinated aliphatic solvents and 1 for aromatic solvents, δ_H^2 is the square of the Hildebrand solubility parameter and accounts for the solvent contribution to create a solute cavity and ξ is a coordinate covalency term, whose value depends on the nature of the basic functional groups present in the solvent. The coefficients a, b, s, h and e indicate the weight of their related solvent parameters on the variation of the XYZ solute property. This extended expression can be reduced by several terms depending on the nature of the solvent-dependent process studied and on the solvents and solutes used in doing so. Actually, a more basic expression which includes only the solvatochromic parameters α, β and π^* is often taken as the starting point in the establishment of LSERs. Linear models constructed from all three parameters or even by only one or two of them have proved useful in the accurate description of a wide variety of solvent-dependent processes (Kamlet,1985).

The three solvatochromic α, β and π^* scales are built on the basis of the LSER philosophy. Thus, the linear model underlying these scales is:

$$\nu_{max} = \nu_o + a\alpha + b\beta + s\pi^*$$

where the spectral shift in the maximum of the absorption band of each reference probe depends on one or more of the terms in the equation above. Thus, the α and β scales work with pairs of reference solutes whose difference in solvatochromic shift correlates with the solvent hydrogen-bond acidity and solvent hydrogen-bond basicity, respectively and the π^* scale is established by using a set of solvatochromic indicators whose spectral shift is only correlated with the solvent polarity/polarizability. All the spectral data used in the determination of α, β and π^* values must be recorded at 25°C to ensure that the spectral shift detected is due only to variations in the solute/solvent interactions and is not caused by thermosolvatochromic effects (Nicolet,1986); the chemical meaning and the accuracy of any solvatochromic parameter determined from spectra recorded at different temperatures would be seriously damaged owing to the variations in the position and shape of the probe absorption band caused by modifications of this chemical variable.

The π^* scale was proposed in 1977 as an empirical measure of solvent polarity/polarizability (Kamlet,1977). In contrast to previous polarity scales of single parameters, the concept of polarity is not understood here as being the solvent overall solvation ability, but as a measure of the solvent effect on the solutes due only to the establishment of pure polarity/polarizability interactions. The specific nature of this empirical parameter depends largely on the solvatochromic indicators selected for its determination. Apart from fulfilling, as much as possible, the general features listed in section 1.1., the choice of these indicators is also focused on reducing the LSER model $\nu_{max} = \nu_o + a\alpha + b\beta + s\pi^*$ used in the general description of any solvent-dependent spectral shift to the more specific expression:

$$\nu_{max} = \nu_o + s\pi^*$$

where the spectral shift is shown to be only affected by solvent polarity contributions. The terms related to hydrogen-bonding drop out from the general LSER model when solvents not able to develop this kind of interactions (NHB) are analyzed. In this case, not much attention should be paid to the probe ability to interact via hydrogen-bonding with the solvent, since $\alpha = \beta = 0$. When dealing with hydrogen-bond acceptor solvents (HBA, $\alpha = 0$ and $\beta \neq 0$) or hydrogen-bond donor solvents (HBD, $\alpha \neq 0$, $\beta = 0$), the reference solutes selected must not present the hydrogen-bonding capacity which complements that of the solvent studied, i.e., they must be non-acidic solutes ($b = 0$) for the HBA solvents and non-basic solutes ($a = 0$) for the HBD solvents. The greatest problem arises when amphiprotic solvents (HBA-D, $\alpha \neq 0$, $\beta \neq 0$) have to be handled because of the difficulty in finding reference probes with no hydrogen-bonding abilities and large enough solvatochromic shifts (i.e., big s values). The best solution here is the combined use of NHB solutes with lower sensitivities and weak HBA probes with higher sensitivities where the hydrogen bonding interactions with the solute are too weak to be noticed or where the autoassociation of solvent molecules is more favourable than the disrupting of this self-association to form hydrogen bonds with the reference solute.

To avoid the presence of the specific effects of one indicator, the π^* values are often determined from the average solvatochromic behaviour of several reference solutes. Forty-five probes classified according to their compatibility with the different kinds of solvent were firstly proposed for application in the determination of the π^* parameter. All these substances present $p \rightarrow \pi^*$ or $\pi \rightarrow \pi^*$ transitions whose ν_{max} are correlated with each other in non-HBD solvents (and sometimes in HBD solvents as well) and whose spectral shifts are clearly consistent with the solvent polarity variation and cannot be attributed to other causes, such as experimental error or spectral anomalies. π^* values equal to 0 for cyclohexane and equal to 1 for dimethylsulfoxide are fixed as reference points on the scale and the terms ν_o and s are calculated for each indicator so that these references are respected. Despite the classification made in the earliest work of Kamlet et al. (Kamlet,1977) concerning the suitability of the indicators proposed for the various kinds of solvent, the results obtained from the several substances used in the determination of π^* values must be carefully checked before being averaged to prevent the appearance of meaningless values which could arise from the inclusion of inadequate indicators showing a behaviour clearly unlike that of the other substances (Cheong,1988). Indeed, several studies have shown that the π^* values are more indicator-dependent than was supposed when this empirical scale was established (Brady,1982) and the recent tendencies are towards the use of only one indicator whenever possible. From the primary set of probes proposed by Kamlet et al., 4-nitroanisole is considered the best reference substance owing to the invariability of its band shape from solvent-to-solvent. When the extreme overlap between the probe and the solvent absorption bands prevents the use of this substance, the N,N-dimethyl-4-nitroaniline (proposed as an alternative to the original probe N,N-diethyl-4-nitroaniline due to its weaker vibrational structure, i.e., to its less solvent-dependent band shape) can be used instead (Laurence,1994).

The α and β scales used to quantify the solvent ability to interact with the solute via hydrogen bonding were established in 1976; the α scale was intended to represent the solvent hydrogen-bond donor acidity (Taft,1976) whereas the β scale accounted for the solvent hydrogen-bond acceptor basicity (Kamlet,1976).

The experimental determination of both α and β parameters is not as straightforward as the procedure used to obtain π^* values. In contrast to the specificity of the polarity indicators, there is no possibility of finding any probe whose spectral shift is related only to hydrogen-bonding interactions with the solvent, i.e., no single substance shows a spectral shift that can be simply defined by LSERs models such as $\nu_{max} = \nu_o + s\beta$ or $\nu_{max} = \nu_o + s\alpha$ because all the solutes capable of developing hydrogen bonding interactions also show polar interactions.

The protocol followed in isolating and quantifying the contributions related to hydrogen-bonding interactions involves the selection of pairs of solutes whose only difference in terms of solvatochromic behaviour is due to their varying ability to establish acidic or basic hydrogen-bonding interactions. These pairs are formed by one substance which is able to develop hydrogen-bonding interactions and another one unable to do so; both substances can be homomorph, such as 4-nitroanisole and 4-nitrophenol used in the determination of β values, or not, such as Reichardt's betaine and 4-nitroanisole used to determine α values. Whatever the nature of the two solutes involved, the pairs of indicators applied in the determination of hydrogen-bonding parameters must fulfil the following three requirements:

a) there must be a linear relationship with a strong correlation between the ν_{max} of both indicators when measured in solvents unable to develop hydrogen-bonding interactions, i.e., modifications in the polar solute/solvent interactions must cause the same kind of variation in the spectral shift of both substances,

b) experimental measurements performed in hydrogen-bonding solvents must show a significant displacement from the aforementioned linear relationship, and

c) the direction and magnitude of the displacements should reflect a reasonable order of solvent hydrogen-bond donor acidity (for the α scale), when one of the solutes in the pair is basic, and of solvent hydrogen-bond acceptor basicity (for the β scale), when one of the reference solutes is acidic.

Bearing in mind these conditions, the α scale is built taking advantage of the enhanced solvatochromic shift for Reichardt's betaine (2) relative to 4-nitroanisole (1) in

HBD solvents. Thus, from the measurements performed in a series of NHB solvents (in kK), the following linear model is found:

$$\nu(2)_{max} = -1.873\nu(1)_{max} + 74.58$$

Experimental $\nu(2)_{max}$ values clearly deviate from the previous model when they are measured in HBD solvents owing to the presence of hydrogen bonding interactions between the acidic solvent and the basic betaine probe which do not take place between the solvent and the 4-nitroanisole. The solvatochromic displacement associated with the hypsochromic shift of the betaine probe in HBD solvents can be quantified as:

$$\Delta\Delta\nu(2\text{-}1) = \nu(2)_{max}(\text{exp}) - \nu(2)_{max}(\text{calc})$$

where $\nu(2)_{max}(\text{exp})$ is the experimental betaine wavenumber in an HBD solvent and $\nu(2)_{max}(\text{calc})$ is the betaine wavenumber calculated for the same HBD solvent by using the linear model relating $\nu(2)_{max}$ and $\nu(1)_{max}$ in NHB solvents. The single fixed reference point of the solvent hydrogen-bonding acidity scale is the methanol α value, set equal to one. Since the methanol solvatochromic displacement, $\Delta\Delta\nu(2\text{-}1)$, has a value of 6.24 kK, the calculation of the hydrogen-bond donor acidity for any other solvent will follow the expression:

$$\alpha_1 = \Delta\Delta\nu(2\text{-}1)/6.24$$

where the subscript 1 in the term α_1 denotes that this solvent parameter has been evaluated by using the solute pair formed by probes 1 (4-nitroanisole) and 2 (Reichardt's betaine). Additional examples of solute pairs proposed to determine the solvent hydrogen-bond donor acidity are Brooker's merocyanine (3) (Brooker,1965) and 4-nitroanisole (1), which yield α_2 values, or Burgess' complex bis[α-(2-pyridylbenzylidene)-3,4-dimethylaniline]bis(cyano)iron(II) (5) (Burgess,1970) and N,N-diethyl-4-nitroaniline (6), used to obtain the α_4 values. Chemical properties which differ from the spectral shifts have also been applied to determine further α_i values.

Analogous to the α scale, the β scale is built using the enhanced solvatochromic shift for 4-nitroaniline (1) relative to N,N-diethyl-4-nitroaniline (2) in HBA solvents. The linear relationship between the ν_{max} (in kK) of both indicators measured in NHB solvents follows the expression:

$$\nu(1)_{max} = 1.035\nu(2)_{max} + 2.64$$

Bathochromic shifts associated with the 4-nitroaniline when dissolved in HBA solvents cause clear deviations in the experimental $\nu(1)_{max}$ values from the previous linear model. These solvatochromic displacements stem from the presence of hydrogen bonding interactions between the basic groups of the solvent and the hydrogen atoms of the amino group in the 4-nitroaniline, interactions which cannot take place with the N,N-diethyl-4-nitroaniline since the hydrogen atoms linked to the amino nitrogen in its homomorph partner are replaced by ethyl groups. The solvatochromic displacement associated with the bathochromic shift of the 4-nitroaniline in HBA solvents is calculated as:

$$-\Delta\Delta\nu(1\text{-}2) = \nu(1)_{max}(calc) - \nu(1)_{max}(exp)$$

where $\nu(1)_{max}(exp)$ is the experimental 4-nitroaniline wavenumber in an HBA solvent and $\nu(1)_{max}(calc)$ is the 4-nitroaniline wavenumber calculated for the same HBD solvent by using the linear model relating $\nu(2)_{max}$ and $\nu(1)_{max}$ in NHB solvents. The single fixed reference point of the solvent hydrogen-bonding basicity scale is the hexamethylphosphoramide β value, set equal to one. Since $-\Delta\Delta\nu(1\text{-}2) = 2.8$ kK for this solvent, the hydrogen-bond donor basicity for any other solvent can be calculated as:

$$\beta_1 = -\Delta\Delta\nu(1\text{-}2)/2.8$$

where the subscript 1 in the term β_1 denotes that this solvent parameter has been evaluated by using the solute pair formed by probes 1 (4-nitroaniline) and 2 (N,N-diethyl-4-nitroaniline). The β_2 values are determined from the enhanced solvatochromic shift for 4-nitrophenol (3) relative to 4-nitroanisole (4) in HBA solvents. Owing to the nature of the solute pairs used to obtain β_1 and β_2 values, β_1 quantifies better the solvent basicity vs. NH donors whereas β_2 characterizes better the solvent basicity vs. OH donors (Laurence,1986). Apart from this last comment, β_1 values are preferred when measuring the basicity of amphiprotic solvents because the spectral shifts of its related solute pair are not so clearly affected as those measured when obtaining the β_2 values by the interactions established between the oxygen atoms in the nitro group of the probes and the acidic groups of the solvent (Kamlet,1976). As in the α scale, β_i values can also be determined by using chemical properties other than spectral shifts, though this alternative is seldom adopted due to the great complexity of the experimental measurements needed.

The differential nature of the experimental measurements used in the calculation of the α and β values, the use of a single reference point in the establishment of both scales and the somewhat solvent family-dependent behaviour of some solute pairs explains the significant scattering of values originating from the different series of α_i and β_i values. The dispersion of these results advises against the use of averaged values which are drawn from diverse α_i or diverse β_i series, as this would lead to a loss of chemical meaning rather than to a balancing of the differences originating from various experimental measurements; the adoption of only one α_i series in defining the solvent acidity and only one β_i series in accounting for the solvent basicity seems to constitute the best alternative for gaining a reliable picture of the true variation of these microscopic parameters in the different solvents.

1.4. Solute behaviour vs. solvent effect: the establishment of Linear Solvation Energy Relationships (LSER).

The goal of all empirical scales of microscopic parameters is their subsequent use in interpreting variations in solvent-dependent solute properties as a function of the modifications of the intensity and the nature of the solute/solvent interactions (Politzer,1994). The relationship between the solute behaviour and the solvent effect can often be expressed with simple linear models, the so-called linear solvation energy relationships (LSER), that present the following general structure:

$$XYZ = (XYZ)_o + \Sigma a_i s_i$$

where XYZ is the solvent-dependent solute property, $(XYZ)_o$ is the value of this property in a hypothetical inert solvent unable to interact with the solute, s_i is a solvent parameter responsible for the variation of the solute property and a_i is the regression coefficient of s_i representing the weight of the s_i contribution to the variation of the solute property XYZ (Kamlet,1981), (Reichardt,1988), (Pytela,1988).

The simplest LSERs are those which correlate a solute property with only one polarity parameter. The success of these uniparametric correlations (many have been reported using the $E_T(30)$ parameter (Reichardt,1982), (Johnson,1986), (Tunuli,1984)) seems to be due to the great similarity between the solvent/probe interactions and the solvent/solute interactions in the particular solvent-dependent process studied. When a single probe does not represent all the necessary solute/solvent interactions, LSERs can be proposed by combining various single parameters whose probes show complementary solute/solvent interactions (i.e., $E_T(30)$, sensitive to polarity and hydrogen bond acidity and the donor number (DN), sensitive to polarity and hydrogen bond basicity (Krigowski,1975), (Wrona,1991)). The only drawback to these combinations is the possible correlation between the parameters used due to the presence of common solute/solvent interactions in some of them (e.g., polarity contributions in $E_T(30)$ and DN).

The most recommendable alternative for the establishment of multiterm LSERs is the use of groups of parameters proposed jointly to cover all the specific solvent/solute interactions (Kamlet,1981), (Swain,1983), (Drago,1992). The advantage of working with parameters belonging to the same original multiparametric approach is that each parameter is designed to have a maximum specificity and a minimum correlation with the other parameters in the same group. This yields linear models with better mathematical features, i.e., with no correlation between terms, and with a clearer chemical sense, i.e., the specificity of each parameter allows the direct interpretation of the contribution of each term to the model as the contribution of its related solute/solvent interaction to the variation of the solute property. The solvatochromic comparison method proposed by Kamlet, Abboud and Taft is the most widely used multiparametric approach for describing the solvent effects on a solute property (Taft,1976), (Kamlet,1976;1977). The LSERs based on this empirical approach are based on the linear model $XYZ = (XYZ)_o + a\alpha + b\beta + s(\pi^* + d\delta) + h\delta_H^2 + e\xi$ (described in section 1.3 above) or on reduced expressions of it, such as $XYZ = (XYZ)_o + a\alpha + b\beta + s\pi^*$. According to the solute property determined and to the solvents used in the study, these expressions can keep all or only some of the terms present in its more extended form (Kamlet,1983;1985), (Taft,1985).

The proven success of the solvatochromic comparison method for explaining the solvent effect in many solute properties has led to the proposal of other general expressions inspired by the same underlying philosophy of tackling the same problem as the original approach or analogous problems where the same kind of formulation is supposed to be adequate. This is the case of the LSERs proposed by Carr (Li,1991), (Carr,1993) and Abraham (Abraham,1993), where solute-to-solute variations of a measured property in a fixed solvent are studied. In this case, the changes in the XYZ property are described using linear models which contain solute descriptors instead of the solvent empirical parameters present in the original expression of Kamlet, Abboud and Taft. The general equation becomes then:

$$XYZ = (XYZ)_o + a\alpha_2 + b\beta_2 + s\pi^*_2 + \sum d_i x_{i2}$$

where the subscript 2 indicates that the independent variables are solute descriptors. α_2, β_2 and π^*_2 are then the solute hydrogen-bond acidity, solute hydrogen-bond basicity and solute polarity-polarizability, respectively, and $\sum d_i x_{i2}$ are other terms in the model including additional solute descriptors which depend on the process analyzed and on the empirical approach applied. $\alpha_{2,}$ β_2 and π^*_2, while sharing the same chemical meaning in both Abraham's and Carr's formulations, have different numerical values owing to the different methodologies employed in determining then (Politzer,1994). Merging the Kamlet, Abboud and Taft basic expression, which works with measures of a property for one solute in a series of solvents, with the last equation, which handles measures of a property for a series of solutes in a fixed solvent, a more general equation with crossed-terms showing the solute and solvent-paired properties involved in each kind of solute/solvent interaction can be obtained.

$$XYZ = (XYZ)_o + a\beta_1\alpha_2 + b\alpha_1\beta_2 + s\pi^*_1\pi^*_2 + \sum d_i\, x_{i1}x_{i2}$$

Subscripts 1 and 2 denote solvent and solute properties, respectively. This equation can be used to describe variations of a property measured for a series of solutes dissolved in a series of solvents (Kamlet,1985).

A recent alternative to all these expressions above are the so-called theoretical linear solvation energy relationships (TLSER), proposed by Famini and Wilson (Famini,1989;1992). These are general models with the same structure and aim as the classical LSERs, in which the empirical solute and/or solvent descriptors are substituted by theoretical descriptors determined from molecular orbital computational methods (Lowrey,1995). Modelled after the experimental LSER parameters, the TLSER descriptors have been designed to correlate closely with their empirical pattern descriptors in order to obtain theoretical models which are as similar as possible to the original LSERs and which can work in a wide range of applications. Within the domain of linear models using theoretical descriptors, Murray and Politzer have developed a rather innovative approach named the general interaction properties function (GIPF) (Murray,1994) (Politzer,1994).

Although it shares the goal of seeking a quantitative linear relationship between a certain property and some microscopic descriptors, GIPF works with theoretical descriptors which are not related to those used in LSERs. The general formulation of the GIPF approach is as follows:

$$\text{Property} = \text{f}[\text{surface area}, \bar{I}_{S,min}, V_{S,max}, V_{S,min}, \Pi, \sigma_{tot}^{2}, \nu]$$

where $I_{S,min}$ reflects the tendency for charge transfer, $V_{S,max}$ and $V_{S,min}$ are indicators of long-range attraction for nucleophiles and electrophiles, respectively, Π accounts for the local polarity, σ_{tot}^{2} for the variability of the surface electrostatic potential and ν is an "electrostatic balance" term. The evolution of these theoretical approaches opens up new possibilities in the interpretation of processes by means of weighted sums of microscopic contributions, such as the determination of solvent or solute descriptors which are not available experimentally and, what is more interesting, the prediction of these descriptors and their related properties for molecules yet to be synthesized, thus contributing to the intelligent design of molecules for special purposes.

In the proposal of any theoretical or empirical LSER, effort must be focused on the correct establishment of the linear model that will afterwards be used to understand the solute/solvent interactions responsible for the variation of a certain property and to predict the values of this property, once the suitable solute or solvent descriptors are known. As pointed out by most of the authors who have proposed general LSER models, not all the terms in these expressions must necessarily be included to describe the variation of the XYZ property of interest (Kamlet,1981), (Reichardt,1988). The existence and weight of the different contributions included in the LSER proposed will depend on the nature of the XYZ property and also on the solutes and solvents employed in its study. At the beginning of the generalized use of LSERs, the procedures applied in establishing these linear models often tended to work with the whole original expressions regardless of the significance of each of their terms for the description of the solute property under study. This rough methodology led some scientists to cast reasonable doubts upon the chemical meaning of these expressions and to consider LSERs as local empirical rules rather than as behaviour models structured as combinations of fundamental microscopic effects (Sjöström,1981). Actually, these early overfitted models do not go beyond the category of local expressions with a limited predictive ability. Nevertheless, owing to the intrinsic quality of their microscopic descriptors and to their sound chemical meaning, the LSERs established using reliable methodologies which check the significance of each of the terms included in the model are valuable expressions that enlighten the nature of the solute or solvent effect on many different properties (Kamlet,1985).

Identifying methods that can ensure the quality of LSERs is one of the concerns of the present project. Two strategies with rather different backgrounds are proposed: the use of hard-modelling methods that fit the experimental data to a postulated chemical model, and the application of soft-modelling methods, where the final model is built without the use of any initial basic expression.

1.4.1. Hard-modelling methods.

All the procedures included within this group share the need of an initial general model to be used in the establishment of the LSER. When the experimental data to be handled are related to the variations of a solute property in a series of solvents, as they are in the case of the examples presented in this project, the reduced expression of Kamlet, Abboud and Taft, $XYZ = (XYZ)_o + a\alpha + b\beta + s\pi^*$, is a highly suitable initial model.

Once the solute property and the solvent descriptors in the initial model have been determined for a certain solvent set, the establishment of the LSER can be carried out. This process includes the following steps:

1. **Selection of the solvent descriptors to be included in the LSER.** A stepwise procedure is used to determine which of the solvent descriptors in the initial model must be present in the LSER. This procedure is a combination of forward selection and backward elimination of variables, as proposed by Forina et al. (Forina,1988). First of all, the solvent descriptor showing the largest correlation with the solute property is selected

as the first independent variable in the LSER. Before the introduction of a new independent variable, two statistical F-tests are computed. In the first, an F-to-enter value is calculated for each non-entered solvent descriptor to check if the introduction of the new descriptor causes a significant decrease in the variance associated with the model fit; in the second, an F-to-delete value, smaller than the F value in the first test, is calculated to see if any of the already selected variables can be removed from the model. This process continues until any of the non-selected descriptors give F values larger than the control F value.

2. **Detection and removal of the outliers present in the data set.** Once the solute property (dependent variable) and the suitable solvent descriptors (independent variables) have been selected, a least median of squares regression (LMS) is applied. This robust approach allows the detection of outliers through the study of the standardized LMS residuals. A function of these residuals is also applied in the robust diagnostic method proposed by Rousseeuw and Leroy (Rousseeuw,1981) for outlier detection. Only data detected as outliers by both previous methods have been removed from the data set in order to avoid the exclusion of good leverage points from the definitive model.

3. **Establishment of the definitive LSER model.** A least-squares fit is performed with the selected solvent descriptors, without including the detected outliers. Despite the precautions taken in the previous steps to ensure the establishment of a correct LSER, a last quality control including a *t*-test to confirm the significance of each of the terms included in the model and an analysis of the variance to support the existence of a correlation between the solute property and the solvent descriptors is performed on the definite LSER expression.

To obtain a clearer idea of the quality of the LSERs established, it is recommendable to provide information regarding the number of points in the data set, the error associated with each of the coefficients in the model and certain parameters referring to the global quality of the correlation, such as the residual standard deviation or the correlation coefficient.

1.4.2. Soft-modelling methods.

These model-free methods work with multivariate data sets and allow the establishment of general LSERs without the need of any initial model. Among the wide variety of soft-modelling procedures, the combination of Factor Analysis (FA) and Target Factor Analysis (TFA) is proposed here given its great ability to describe the variation of multivariate data sets through the establishment of meaningful linear models (Malinowski, 1991).

The experimental data required in understanding the solvent effect on a solute property are measurements of the property under study in a certain solvent set. Several solutes are often used to record these series of measurements in order to avoid biased conclusions about the solvent effect which could arise if the research was performed with only one solute owing to its concrete features. The measurements obtained for the different solutes in the selected solvent set can be structured to form a data matrix, as shown in Figure 1.4.2.1.

$$\begin{array}{c|ccc|}
 & \text{Solute 1} & \text{Solute 2} & \text{Solute n} \\
\hline
\text{Solvent 1} & a_{11} & a_{12} & a_{1n} \\
\text{Solvent 2} & a_{21} & a_{22} & a_{2n} \\
 & & & \\
\text{Solvent m} & a_{m1} & a_{m2} & a_{mn} \\
\hline
\end{array}$$

Figure 1.4.2.1. Arrangement of solvent-dependent arrays of data to form a data matrix. a_{ij} represents the value of an experimental property for the solute j dissolved in the solvent i.

Each column in the data matrix is the array of data collected for a certain solute; hence, a solvent-dependent variation is present along the columns of the data matrix whereas solute-to-solute differences are responsible for the variation detected along the rows.

Factor Analysis is a frequently used technique for interpreting the underlying causes of variation in the data matrices. This chemometrical procedure decomposes the original data matrix **D** *(r × c)* into the product of the scores matrix **T** *(r × n)* and the loadings matrix $\mathbf{P}^T$ *(n × c)*, whose column vectors and row vectors describe the *n* independent abstract sources of variation (factors) along the columns and along the rows of the original data matrix, respectively. This matrix decomposition can be rewritten as a series of additive contributions formed by the outer products of each score vector $\mathbf{t}_i$ by its related loading vector $\mathbf{p}_i^T$.

$$\mathbf{D} = \mathbf{T}\mathbf{P}^T$$

$$\mathbf{D} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_n\mathbf{p}_n^T$$

Hence, the application of FA only makes sense if the original data matrix is intrinsically bilinear, i.e., if it can be described with a model of additive contributions related to the real sources of variation in the data.

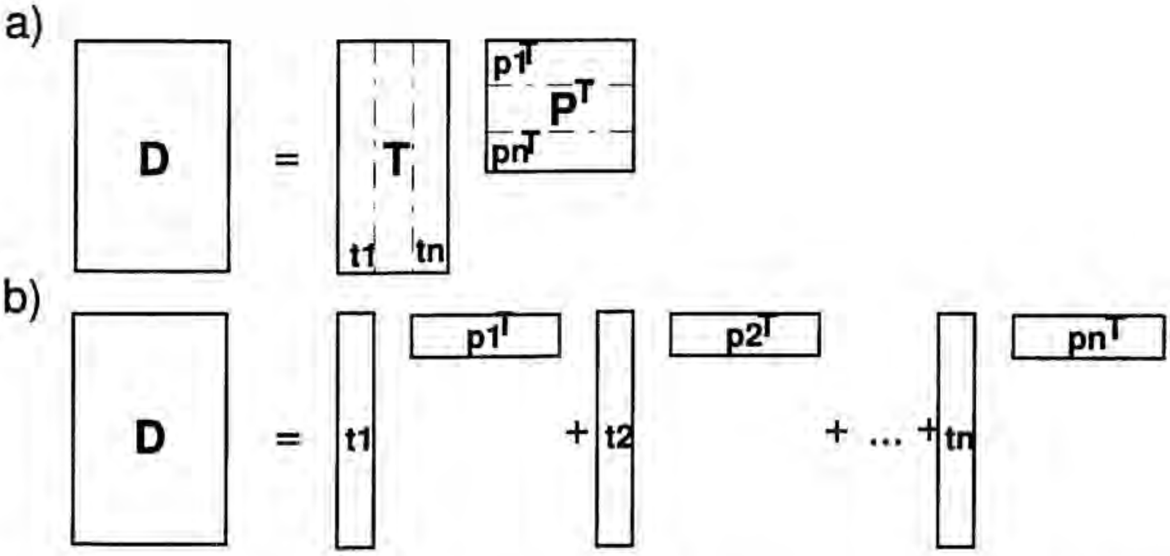


Figure 1.4.2.2. FA decomposition of a data matrix as a) product of scores and loadings matrices and b) as a sum of the outer products of related scores and loading vectors.

The scores and loading matrices are calculated from the diagonalization of the covariance matrix $\mathbf{Z} = \mathbf{D}^T\mathbf{D}$ associated with the **D** matrix. This process is carried out by finding a **Q** matrix such that

$$\mathbf{Q}^{-1}\mathbf{Z}\mathbf{Q} = \lambda$$

The columns of **Q** are the eigenvectors of the **Z** matrix and the elements in the λ matrix are their related eigenvalues. The magnitude of each eigenvalue is linked to the importance of its related eigenvector when describing the variation of the **D** matrix. $\mathbf{Q}^T$ is equal to the loadings matrix, $\mathbf{P}^T$, and owing to the orthogonality of the **Q** matrix, the scores matrix **T** can be simply calculated as

$$\mathbf{T} = \mathbf{D}\mathbf{Q}$$

Due to the process followed in the FA decomposition of a data matrix, there is always a lack of correlation among the column vectors in the scores matrix **T** and among the row vectors in the loadings matrix $\mathbf{P}^T$.

When a data matrix such as that shown in Figure 1.4.2.1. is factor-analyzed, the score vectors represent independent solvent contributions to the overall variation of the data matrix. These vectors are actually abstract solvent descriptors because they depend on the solvent features but cannot be directly identified with any chemically meaningful solvent parameter. The elements in a loading vector are the weights of their related abstract solvent descriptor for each of the solutes in the original data set. The original data matrix (i.e., the variations of a solvent-dependent property measured in a series of solutes) is then defined by using the simplest model of additive contributions formed by the outer products of an abstract solvent descriptor vector by a vector containing the solute-weighted contributions of this abstract descriptor on the variation of the property under study.

The FA decomposition of a data matrix recalls the structure of an LSER model; indeed, by looking at the Figure 1.4.2.3. it can be seen that both ways of expressing the original data matrix share exactly the same form. FA and LSER describe the original data matrix as a sum of solute-weighted contributions of solvent descriptors (abstract in the case of FA and chemically meaningful in the LSER model) or, in a more compact way, as the product of a matrix of solvent descriptors by a matrix containing their related weight coefficients to explain the overall variation of the data matrix.

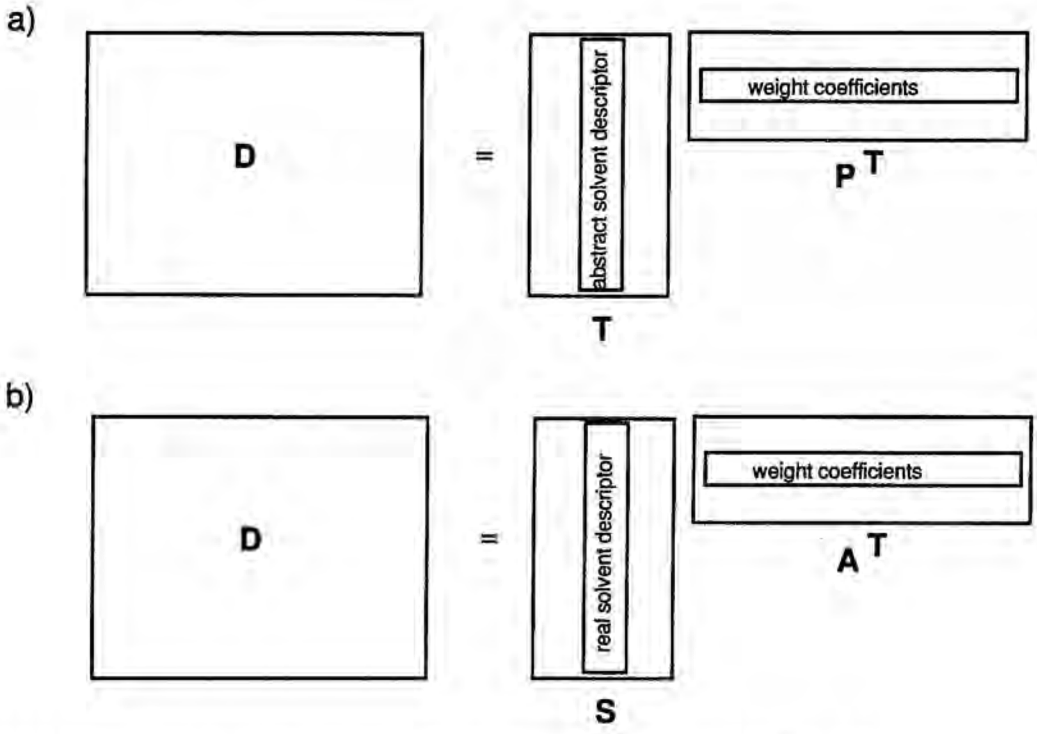


Figure 1.4.2.3. Data matrix description by using the FA decomposition or an LSER model.

Though FA describes perfectly the variation of the experimental data with an abstract model, the ideal situation would be to find a procedure by which the abstract solvent descriptors could be transformed into chemically meaningful solvent parameters. Target Factor Analysis (TFA) is a technique whose main goal is determining the connection between the domain of abstract solutions and the domain of real solutions (exemples,).

The combined application of FA and TFA to establish an LSER model includes the steps listed below, and which are later explained in detail.

1. Application of FA in determining the number of significant contributions to be included in the LSER model.
2. Selection of the targets proposed as potential solvent descriptors in the LSER model.
3. Target testing.
4. Building the definitive LSER model by using combinations of accepted targets.

1. **Application of FA in determining the number of significant contributions to be included in the LSER model.** In the FA decomposition of a data matrix, the number of eigenvectors calculated is equal to the number of rows or the number of columns of the matrix, whichever is the smallest. The eigenvectors are then decorrelated and ranked in descending order according to their ability to account for the variation present in the data. Thus, the first eigenvector (factor) lies in the direction of the greatest percentage of variation in the data, the second in the direction related to the next greatest percentage and so on, till the complete variation of the data, including the noise contributions, is explained. Once all these eigenvectors have been calculated, the first point must be focused on seeking how many of them represent chemically meaningful sources of variation. The number of significant factors *n*, i.e., the true rank of the matrix, will indicate the number of terms to be included in the future LSER model.

Several methods with different backgrounds have been used to determine the number of significant factors in the data matrix. Based on the theory of error in FA proposed by Malinowski, the correct number of factors can be found by looking at the minima in the real error (RE) function and in the IND function (Malinowski,1987). These two functions are determined depending on the number of factors considering *n*, as follows:

$$\mathrm{RE} = \sqrt{\frac{\sum_{j=n+1}^{c} \lambda_j}{r\,(c-n)}}$$

$$\mathrm{IND} = \frac{\mathrm{RE}}{(c - n)^2}$$

where λ_j is the eigenvalue associated with the jth eigenvector, *c* is the total number of eigenvectors calculated, i.e., the number of columns in the data matrix if this is the smallest dimension in the matrix and *r* is the number of rows in the data matrix. Cross validation techniques, as proposed by Wold (Wold,1978) and Malinowski (Malinowski,1987), have also been applied; in this case, minima in the evolution of the standard error in prediction (SEP) function indicate the correct number of factors. With a greater statistical basis than the previous procedures, an F-test (Malinowski,1988) has also been performed to distinguish the eigenvalues related to noise from those related to significant factors. All these methods operate rather well when handling matrices with a randomly distributed error. If the error in the data matrix shows a certain structure, e.g., heteroscedasticity, drift,..., the procedures above tend to overestimate the number of significant factors in the data set.

2. **Selection of the targets proposed as potential solvent descriptors in the LSER model.** This is the essential part of the FA-TFA process. A target vector is a potential real factor that can be used to explain the variation in the data matrix analyzed. Therefore, the proposal of targets obeys scientific reasoning and never comes from a random vector selection. The sounder the knowledge is about the chemical problem under study, the greater are the possibilities to select appropriate targets to build the real model related to the variation of the data set.

   In the establishment of an LSER model, the targets used must be potential solvent descriptors, e.g., $E_T(30)$, α, β, π^*, $1/\varepsilon$, ... There is no limitation to the number of targets which can be proposed and, in contrast to the hard-modelling procedures, the solvent descriptors should not come necessarily from only one uniparametric or multiparametric approach. An additional target must always be proposed in the establishment of any kind of linear model: the unity target. This vector, whose elements are all equal to one, accounts for the possible need to include a constant contribution when explaining the variation in each column of the original data matrix or, in other words, this target considers the possibility of introducing an offset in the LSER model.

3. **Target testing.** This operation assesses the usefulness of the proposed targets in building the LSER model. The previous application of FA to the data matrix yielded *n* significant score vectors which span perfectly the data space. A real factor able to describe the variation of the data must necessarily belong to this space. The value of a target will therefore be assessed by testing whether or not the target lies in the space defined by the score vectors.

   From a geometrical point of view, the operation carried out to test each selected target individually is the projection of this input target, $\mathbf{s}_i$, onto the scores space. The projected vector, the so-called output target, $\mathbf{s}_o$, belongs to the scores space. Figure 1.4.2.4. shows the target testing procedure for a two-factor system; i.e., an example where the data space can be graphically represented with a plane.

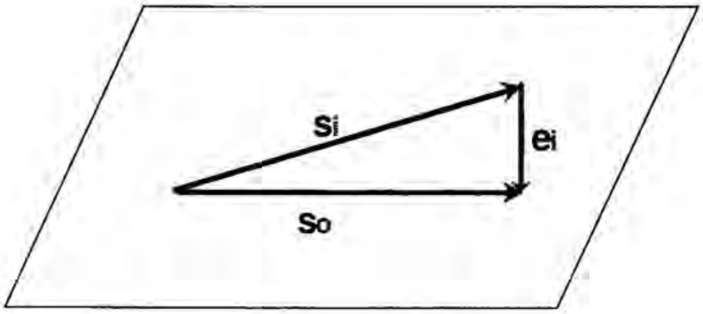


Figure 1.4.2.4. Target testing procedure on a data space defined by two factors.

Mathematically, the target testing procedure begins with the calculation of a transformation vector, **y**, which explains the relationship between the scores space, represented by the scores matrix, **T**, and the input target, $\mathbf{s}_i$.

$$\mathbf{s}_i = \mathbf{T}\mathbf{y}$$

$$\mathbf{y} = \mathbf{T}^{+}\,\mathbf{s}_i$$

where $\mathbf{T}^+$ is the pseudoinverse of the **T** matrix. Once y is known, the output target can be calculated as follows:

$$\mathbf{s}_o = \mathbf{Ty}$$

A perfect target would be a vector for which $\mathbf{s}_l = \mathbf{s}_o$. Owing to the experimental error in the data matrix and in the targets proposed, the strict fulfilment of the aforementioned equality is practically impossible. In practice, a target is accepted if the $\mathbf{s}_l$ and $\mathbf{s}_o$ vectors are close enough to each other, i.e., if the length of the apparent error in the target, $\mathbf{e}_l = \mathbf{s}_o - \mathbf{s}_l$, is small enough.

Comparison between analogous elements in the input target and in the output target provides a rough idea about the quality of the target. However, the acceptance or rejection of a target must be supported on additional criteria other than this qualitative observation. Most of the procedures applied are based on the theory of error in Target Factor Analysis proposed by Malinowski (Malinowski, 1991). This author decomposes the error in a target in three contributions related in a Pythagorean way, as shown in Figure 1.4.2.5.

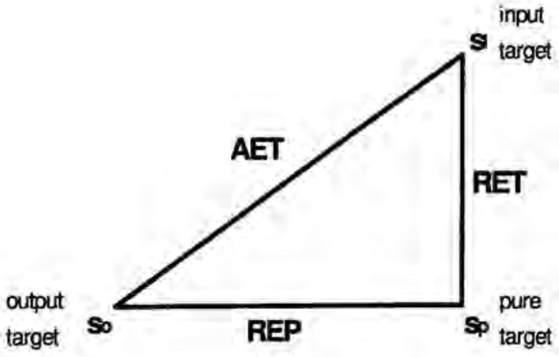


Figure 1.4.2.5. Errors associated with the target testing procedure.

AET is the length of the apparent error in the target vector, $\mathbf{e}_l$, RET describes the real error in the input target and REP the error associated with the output target. By using combinations of these errors, two different criteria in deciding whether to accept a target are proposed. The SPOIL function (Malinowski,1978) has an empirical origin and is calculated as follows:

$$\mathrm{SPOIL} = \mathrm{RET}/\mathrm{EDM}$$

where EDM is the experimental error in the original data matrix, often assumed to be equal to REP. As a rule of thumb drawn from the study of reference sets of data, Malinowski recommends the acceptance of targets whose SPOIL value lies between 0 and 3. Larger values of SPOIL indicate that the use of the target tested would introduce too large errors in the data matrix reproduction. A statistical F-test that compares the magnitudes of the apparent error in the target and the experimental error in the original data matrix is also used (Malinowski,1988). F is calculated as follows:

$$F = \mathrm{AET}/\mathrm{EDM}$$

If the calculated F value is significantly bigger than the theoretical F, i.e., if the uncertainty associated with the reproduction of the target is larger than the experimental error of the original data, the target is rejected.

4. **Building the definitive LSER model by using combinations of accepted targets.** After target testing, several solvent descriptors have been shown to be suitable candidates for inclusion in the LSER model. A number of accepted targets equal to the number of significant factors must be selected. If the number of accepted targets exceeds the number needed to build the LSER model, the targets chosen will be those with the lowest SPOIL and calculated F values. The last part of the TFA process consists of confirming the validity of the LSER model built from the selected real solvent descriptors. All the real solvent descriptors form the **S** matrix, equal in size to the scores matrix. The matrix with the coefficients of these solvent descriptors, **A**, can be calculated taking into account the close relationship between the abstract FA model and the LSER model. The relationship between the matrix having the real targets, **S**, and the scores matrix, **T**, can be expressed as follows:

$$\mathbf{S} = \mathbf{TY}$$

where the columns of the Y matrix are the y transformation vectors related to each of the accepted targets. Since both the FA and the LSER model describe the same original matrix,

$$\mathbf{TP}^{\mathrm{T}} = \mathbf{SA}^{\mathrm{T}}$$
$$\mathbf{TP}^{\mathrm{T}} = \mathbf{TYA}^{\mathrm{T}}$$

must be fulfilled and hence:

$$\mathbf{A}^{\mathrm{T}} = \mathbf{Y}^{-1}\mathbf{P}^{\mathrm{T}}$$

The reproduction of the original data matrix by means of the LSER model is then carried out as shown below:

$$\mathbf{D}^* = \mathbf{SA}^{\mathrm{T}}$$

The root mean square error (RMS) is a measure of the quality in the reproduction of the original data matrix and is expressed with the following equation.

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(d_{ij} - d_{ij}^{*}\right)^{2}}{r \times c}}$$

where d_{ij} is the ijth element in the original data matrix, d_{ij}* is the ijth element of the data matrix reproduced with the LSER model, **D***, and $r \times c$ are the number of elements in the data matrix.

If the RMS error associated with the reproduction of the original data matrix is within the range of the experimental error, the LSER model can be accepted. If this is not the case, a possible correlation exists between some of the targets included in the model and then a different combination of accepted targets must be used to form a new LSER model which can be shown to be valid.

CHAPTER 2.

SOLVENT MIXTURES.

2.1. Seeking the perfect solvent: the alternative of solvent mixtures.

Extensive research into pure solvents and their applications has been carried out to help chemists select the most suitable solvent for their chemical problems. Nevertheless, the ideal choice in most situations would be a medium sharing the features of several pure solvents rather than one specific pure solvent. This solvent, tailored according to the needs of the user, might actually exist among the large family of solvent mixtures.

Despite the evident potential of solvent mixtures, the complex behaviour of these media, where many solvent/solvent and solute/solvent interactions occur, often discourages the non-expert from using them for practical purposes. Indeed, what is generally known is that no single property of a solvent mixture can simply be determined from the composition-weighted sum of the values that the same property has in the solvent partners of the mixture or, in other words, that the behaviour of a solvent mixture is seldom linear with respect to its composition.

Thermodynamic studies and theoretical models have described the internal structure of solvent mixtures and their effect on solutes (Covington, 1974; 1976; 1989), (Marcus, 1989; 1990). However, these rigorous approaches are quite complicated in formulation and their application to the study of complex solute processes in solvent mixtures is either impossible or extremely difficult in practice. Therefore, the use of empirical methods which characterize the solvent mixtures might offer a good alternative which, while less rigorous in formulation, might allow a wider variety of chemical problems to be tackled.

The need for an operational process for the description of solvent mixtures immediately suggests the use of the microscopic parameters mentioned in chapter 1. These parameters would give average values of the solvent mixture ability to develop various kinds of interactions with the solutes (e.g. polarity, hydrogen-bonding). Obviously, no information about the identity of the species performing the interaction with the solute in the solvent mixture (i.e. solvent *i*, solvent *j* or a complex formed by both *i* and *j* solvents) could be obtained, but this information does not necessarily have to be known to interpret the nature and the extent of the overall solvent effect on the solute processes.

The number and variety of solvent mixtures characterized by means of empirical parameters is growing continuously and there are many successful examples of LSERs which describe the effects developed by solvent mixtures on solute processes, such as those proposed in explaining chromatographic retention mechanisms (Carr,1993) (Rosés,1993) (Li,1991). Nevertheless, the validity of the empirical parameters for characterizing solvent mixtures is still questioned because of the solute-dependent nature of the measurements used in determining them. As mentioned in chapter 1, all these empirical scales inherently assume that the solvent behaviour around the probe molecule is the same as around the molecule of any other solute. If this statement could be reasonably accepted for pure solvents, typical phenomena linked to solvent mixtures, such as the preferential solvation (i.e. the existence of differences between the mixture composition in the bulk solvent and around the solute molecules), give rise to doubts about the existence of one common behaviour for a solvent mixture around all solutes.

The first reported use of the $E_T(30)$ scale to characterize solvent mixtures dates from the sixties (Dimroth,1963) and the publication of new values for some other solvent mixtures and the revision of earlier determinations have not stopped since then (Maksimovic,1974), (Koppel,1983a;1983b), (Dawber,1990), (Bosch,1992), (Drago,1994). In addition to the research purely focused purely on the characterization of mixtures, there has been a constant interest in relating the variation of the $E_T(30)$ values and the composition of the solvent mixtures. Pioneering work was carried out in this field by Langhals (Langhals,1982), who proposed the first meaningful equation that expressed the variation of the $E_T(30)$ values of diverse binary solvent mixtures as a function of certain chemical parameters. Langhals' model for binary solvent mixtures is expressed as:

$$E_T(30) = E_D \ln\left(\frac{c_p}{c^*} + 1\right) + E_T^o(30)$$

where c_p is the molar concentration of the most polar component in the mixture, $E_T^o(30)$ is the $E_T(30)$ value for the least polar component and E_D and c^* are the adjustable parameters in the model. E_D is a measure of the sensitivity of the $E_T(30)$ scale towards changes of c_p, whereas c^* is the concentration value that indicates the change in the relationship between $E_T(30)$ and c_p. Thus, when $c_p << c^*$, a linear relationship between $E_T(30)$ and c_p results whereas, when $c_p >> c^*$, the linear relationship is established between $E_T(30)$ and $\ln c_{p.}$ The c^* parameter separates the zones of linear and logarithmic dependence between $E_T(30)$ and c_p and according to Langhals indicates the threshold value at which the two solvents in the mixture begin to interact with each other.

The non-additive behaviour of solvent mixtures can be clearly seen through the evolution of $E_T(30)$ values with solvent composition. When the components of a solvent mixture behave as they would do as pure solvents, the following model can be expected:

$$E_T(30)_{\text{mixture}} = \sum_i x_i \, E_T^o(30)_i$$

where $E_T^o(30)_i$ is the value of this parameter for the pure solvent *i* and x_i is its molar fraction in the mixture. The fulfilment of this equation seldom occurs and deviations from the ideal behaviour are indicative of the preferential solvation of the probe (Haak,1986). The pattern and intensity of these deviations are as diverse as the solvent mixtures themselves and this can be clearly seen by looking at the examples provided by the binary mixtures, the simplest and most widely used mixed media. Apart from methanol-ethanol (Koppel,1983a) or 1,2-dibromoethane-1,2-dibromopropane (Balakrishnan,1981), which belong to the small group of ideal binary mixtures because of the absence or the very weak specific interactions between the two solvent partners in the mixture, the remaining mixtures can be classified in different groups according to their particular kind of preferential solvation on the betaine dye. Thus, in some mixtures one of the solvents is preferentially involved in the solvation within the whole composition range, either the most polar (e.g. dimethylsulphoxide-acetone, ethanol-acetone) or the least polar (e.g. water-methanol), whereas in other mixtures, the preferred solvent changes according to the mixture composition (e.g. in water-acetone mixtures, mixtures with low water mole fractions show water preferential solvation whereas mixtures with high water mole fractions show acetone preferential solvation) (Koppel,1983a), (Dawber,1988). The deviations from the ideal behaviour in mixtures also show marked differences in intensity. Thus, as a general rule, hydroorganic mixtures show weaker deviations than mixtures with two non-aqueous components (Dawber,1983), (Marcus,1994a). In the latter group, amazing synergetic phenomena can even take place, i.e. the mixture can have $E_T(30)$ values higher than those of the two pure solvents (e.g. acetonitrile-ethanol, dimethylsulphoxide-*tert*-butanol); these unusual effects are normally due to the formation of complexes between the solvent partners which have a solvation power which is stronger than the pure solvents (Koppel,1983b). The easiest way to visualize the presence and pattern of the preferential solvation of Reichardt's betaine using a binary mixture consists of displaying the $E_T(30)$ experimental values vs. the bulk molar fraction of one of the components (x_i) in the mixture.

Recent efforts in the field of the preferential solvation have focused on the search for a model that can explain deviations from the ideal behaviour of the solvent mixtures on a chemical basis. In this sense, the positive contribution of works including the equilibria of solvent-solvent complex formation in the formulation of the preferential solvation model must be stressed (Bosch,1995), (Skwierczynski,1994). Nevertheless, the most important point to all this research is the fact that the evolution of this empirical parameter with the mixture composition is very similar to the variation of other kinetic and thermodynamic solute processes (Koppel,1983b) (Dawber,1988) and that the composition of the mixture around the probe is reasonably similar to the composition predicted when using other theoretical approaches. These agreements confirm the finding that the preferential solvation around the probe can be considered as a good model for a large number of solutes and that the $E_T(30)$ values have chemical meaning when determined in solvent mixtures.

Similar studies have been conducted to check the validity of α, β and π* parameters in solvent mixtures. Though no universal conclusions can be drawn, it seems that for a large number of solvent mixtures these parameters hold the same chemical meaning as they hold in pure solvents. The extent of the probe-dependence of these empirical parameters can easily be seen when they are determined as averaged values drawn from several probes. In this case, if the preferential solvation changes from probe to probe, a large scattering in the results from the different probes will be noticed. The determination of the π* parameter for several solvent mixtures has shown that the spread in the π* values derived for the individual indicators in the mixtures is not significantly larger than the similar spread encountered for pure solvents. Similar results have been found for the β parameter, whereas the spread in α values determined by different probes seems to be somewhat larger (Migron,1991).

A different treatment is given to the hydroorganic mixtures and to the completely non-aqueous mixtures (Marcus, 1994a; 1994b). The former group usually presents weak phenomena of preferential solvation and, therefore, the use of empirical parameters is less questionable. Non-aqueous mixtures are more strongly affected by the probe-dependent nature of the empirical parameters, but even in this case, through the selection of probes similar to the solutes involved in the processes where the solvent effect is to be assessed, the empirical approach is still applicable.

2.2. The water-dioxane example.

Water-dioxane can be considered one of the most singular hydroorganic mixtures because of the highly distinctive features of its solvent partners. Bearing in mind that water is an amphiprotic solvent with a high relative permittivity, $\varepsilon = 78.5$, and dioxane is an aprotic ether with a very low one, $\varepsilon = 2.2$, this is likely to be one of the few examples where solvents which are so different are miscible at any ratio. This particularity provides the chemist with a set of solvent mixtures covering a wide range of chemical features depending on the different proportion of the two components in the mixture, as can be seen in table 1.

Table 2.2.1. Some macroscopic properties of water-dioxane mixtures.

% dioxane (v/v)	Molar fraction	Relative permittivity
0	0.000	78.5
10	0.023	70.3
20	0.050	61.4
30	0.083	52.6
40	0.123	44.3
50	0.174	35.7
60	0.241	27.5
70	0.330	19.3
80	0.458	13.5
90	0.655	6.1
100	1.000	2.2

The popularity of this solvent mixture dates from the early forties, when Calvin and Wilson proposed a systematic working procedure to apply to these solvent mixtures in the study of complex equilibria (Calvin,1945). The pioneering work of Van Uitert and Haas (Van Uitert,1952) should be mentioned due to their efforts invested in proposing procedures for measuring the real pH values in these mixtures and in determining the relationship between these measurements and those obtained in the aqueous pH scale. Since then, water/dioxane mixed solvents have been applied to the study of diverse equilibrium processes because of the solubilizing power of these mixtures for many compounds insoluble in aqueous solution. The great variation of properties shown by these solvent mixtures is also very helpful in the interpretation of the solvent effect in many reaction processes.

In reference to the internal structure of water/dioxane mixtures, some research has been conducted which has focused on the characterization of these mixtures by measuring their macroscopic properties, such as relative permittivities (Åkerlöf,1936), (Critchfield,1953), (Mui,1974), while others have studied the solvent interactions between both partners in the mixture. The latter studies propose a change in the internal structure of the water/dioxane mixtures as the proportion of organic cosolvent increases. Thus, there would be an initial water-rich group of mixtures where the typical three-dimensional structure formed by the water molecules would exist and where the dioxane molecules would fill the cavities left by this water network, and a second dioxane-rich group of mixtures, where the structure of the water molecules would be broken down giving as a result free water molecules and water-dioxane complexes of varied compositions formed via hydrogen-bonding between the hydrogen atoms of water and the oxygen atoms of dioxane (Langhals,1982), (Burger,1983). The composition proposed for these water-dioxane complexes ranges from 1:1 to 1:2 or 2:1 in most cases (Arnett,1962), (Glover,1965), (Mohr,1965), depending on the composition of the mixture. The breakdown of the water structure supports the explanation of the changes observed in some classical studies related to the specific solvation associated with water/dioxane mixtures (Erdey-Grúz,1973). Thus, a preferential solvation performed by dioxane is found in water-rich mixtures because the "free" cosolvent molecules can have access to the solute more easily than the network water molecules, whereas this behaviour is modified in dioxane-rich mixtures, where the water molecules released from the network structure are preferentially placed around the solutes.

An additional quality of dioxane/water mixtures is the adequacy of dioxane in a wide variety of experimental conditions and using various instrumental techniques. Indeed, dioxane has certain properties which are quite similar to those of water, such as its boiling point, T_{eb} = 101 °C or its density, $\rho = 1.1$, and therefore, it does not narrow the range of experimental working conditions of aqueous solutions. Furthermore, it does not cause problems in the recording of e.m.f. readings as many organic solvents do and it performs fairly well in spectrometric techniques, such as UV, where it does not absorb at $\lambda > 220$ nm, or circular dichroism where, owing to its symmetrical structure, it does not absorb at all.

2.3. Acid-base equilibria in water/dioxane mixtures.

Acid-base equilibria in water/dioxane mixtures can be formulated as they are in aqueous solution. Dioxane is an aprotic solvent and, therefore, the autoprotolysis of their hydroorganic mixtures occurs only because of the amphiprotic character of water. The presence of dioxane in the hydroorganic mixture is noticed by the hindrance of the water dissociation tendency, i.e. the decrease in the autoprotolysis constant (K_s), and by the shifts induced in other equilibria, but this does not cause the appearance of any additional equilibrium process governed solely by it (Glover,1965). However, this does not mean that dioxane is present in solution as a passive isolated compound. As mentioned before, the abnormal hydrogen-bond basicity of dioxane in hydroorganic mixtures caused by the cooperative influence of the oxygen atoms of dioxane in their interaction with water molecules promotes the formation of dioxane-water complexes with different stoichiometries depending on the composition of the solvent mixture. So, strictly, the autoprotolysis equilibrium of this hydroorganic mixture and the acid-base processes developed in it should be written as

$$2SH \leftrightarrow S^- + SH_2^+$$

$$SH + HA \leftrightarrow SH_2^+ + A^-$$

respectively, where SH is the generic representation of the amphiprotic species in the hydroorganic mixture, i.e. water or any complex dioxane-water and HA is an acidic solute dissolved in the solvent mixture. As can be seen, this notation is the same as that used in aqueous solution if SH is substituted by H_2O. The lack of complexity in the formulation of equilibria in this solvent mixture is an important feature to be taken into account since the chemical behaviour of systems as complex as those studied in aqueous solution can also be understood in media with very different features.

In this project, the study of all the acid-base equilibria was carried out at a constant ionic strength. This procedure, which has been used by many authors when working with these solvent mixtures (Mui,1974), (Das,1980), (Sigel,1993), greatly simplifies the calculation of the equilibrium constants since the knowledge of the activity coefficients related to the species involved in the equilibria is not necessary. Despite possible objections to this procedure when dealing with solvent mixtures with high dioxane contents, because of the possible presence of ion pairs formed by the background electrolyte used in order to keep the ionic strength constant, doubts may be cast upon the existence of such species in the solutions used in the present work because of the high water content in the mixtures with the highest dioxane proportion (70% (v/v) dioxane, water mole fraction = 0.66). These associated species, if present, would probably not occur in a proportion large enough to affect significantly the evaluation of the constants related to the equilibria that take place in the solvent mixture.

2.3.1. Potentiometric studies.

The great accuracy of the potentiometric technique in the determination of equilibrium constants and the nernstian behaviour of most of the electrodic systems in water dioxane mixtures led to the selection of this instrumental technique for studying the autoprotolysis equilibria of water-dioxane solutions covering dioxane concentrations from 10 to 70 % (v/v) and the acid-base behaviour of several simple substances dissolved in these solvent mixtures.

There are two problems inherently associated with the pH-related e.m.f. readings performed in solvents other than water: the increase of the liquid junction potential due to the different nature of the solvent in the working solution and of the solvent in the inner solution of the electrodic system and the calibration of the electrodic system due to the scarce pH standard solutions in the solvent used (Mussini,1984).

In this project, the two difficulties mentioned were solved or at least greatly diminished by applying the following measures to all the experiments carried out. The minimization of the liquid junction potential was achieved by using the following potentiometric cell:

G.E. / working solution, n% dioxane / R.E. (KCl_{sat}, n% dioxane)

where G.E. is the glass electrode, R.E. is the reference electrode and n is the percentage (v/v) of dioxane in the solvent mixture used in the working solution and also in the inner solution of the reference electrode. Problems in the standardization of the electrodic system have classically been solved with the application of the correction functions proposed by Van Uitert and Haas (Van Uitert,1953), which relate the pH readings performed in water-dioxane mixtures with a potentiometer calibrated with aqueous standard solutions to the actual hydrogen ion concentration in these solvent mixtures. Despite the wide application of this approach, all these correction functions are dependent on the nature and concentration of the background electrolyte and the acid used in their establishment and a reappraisal of them is often required according to the working conditions adopted by the chemist. For the sake of simplicity, the use of the Gran method (Gran,1952), an *in situ* calibration process which gives an individual value of electrodic standard potential for each experiment without the need of external standard solutions, was preferred in this project.

All the potentiometric titrations were performed at 25 °C, keeping a constant value of ionic strength. The automation of the potentiometric setup allowed a better control of the experimental conditions (titrant additions, stability criterion associated with the e.m.f. readings,...) and the storage of the experimental data for their subsequent processing.

2.3.2. Data treatment

The acidity constants related to the solutes analyzed were calculated by treating all the experiments related to the same solute together in a given water-dioxane mixture with a least-squares curve fitting procedure based on the iterative refinement of a postulated chemical model. This data treatment was applied as implemented in the SUPERQUAD program by Gans et al. (Gans,1985). This program was specially designed to deal with e.m.f. or pH data related to acid-base or complexation equilibria and determines the constants associated with these equilibria by minimisation of an error-square sum based on the experimental measurements.

SUPERQUAD can be applied to experimental data and chemical systems that fulfil each of the following requirements:

- **possible postulation of a meaningful chemical model.** A chemical model including the equilibria that are likely to take place during the titration must be proposed. This model must define the different reaction processes occurring in solution giving the stoichiometric coefficients of the species involved and an estimation of the value of their equilibrium constants.
- **compliance of the mass action law in the occurring equilibria.** In all stages of a certain equilibrium, the value of the related constant must remain invariant. This discards the use of this data treatment for certain macromolecular systems or for experiments where variations in the equilibrium constant are caused by changes in the solution temperature, ...
- **fulfilment of the mass balance at all the titration points.** The program does not take into account precipitation phenomena.

- **constant ionic strength throughout the titration process.** The equilibrium constants are calculated as concentration ratios; therefore, the quotient of activity coefficients must be constant for all the titration points.
- **nernstian response of the electrodic system.**
- **absence of systematic errors in the data.** Though total certainty can never be guaranteed in the fulfilment of this requirement, the chemist should seek to avoid errors which can be easily prevented, such as those related to a careless experimental work.

The input needed to run the SUPERQUAD program must include:

- the total concentration of each analyte involved in the titration process, i.e., the sum of the concentration values of all its derived chemical species,
- the standard potential of the electrodic system,
- the experimental pairs (titrant volume, measured potential) at each titration point,
- the uncertainties associated with the measures of titrant volume added (σ_V^2) and with the e.m.f. readings (σ_E^2) and,
- a chemical model including all the equilibria that are supposed to occur during the titration with an estimate of their related equilibrium constants.

At each titration point, the error in the measured potential is calculated and the weight assigned by default to each point is inversely proportional to the calculated variance at that point. This weighting procedure gives less importance to the data measured in the extremes of the equilibria processes in which the calculated species concentrations are known to be affected by higher errors. Other weighting procedures can be opted for.

The least-squares algorithm always takes the titrant volume as the independent variable and the e.m.f. measured as the dependent variable. The parameters to be iteratively optimized are normally the equilibrium constants (all or only some) related to the chemical system, though some other parameters, such as the standard potential or various analyte concentrations can also be subject to refinement. Nevertheless, as usual in all iterative procedures, the user is warned against the optimization of too many parameters which might provide results with excellent fits, which are in reality nothing other than mathematical artefacts lacking chemical sense.

In each cycle of the optimization process, new values for the parameters which are refined are evaluated and an estimation of the error associated with the results is given by the parameters σ, which is the ratio between the root mean square of the weighted residuals and the estimated error calculated from the uncertainties σ_V^2 and σ_E^2, and χ^2, which gives information related to the distribution of the e.m.f. weighted residuals. The validity of the chemical model postulated is also assessed in each iterative cycle and all the ill-defined constants (taken as those with negative values or those with relative standard deviations bigger than 33% of its value) are removed from the model in next iterations. The iterative process is finished when convergence is achieved or when a certain number of iterative cycles is exceeded.

The determination of the autoprotolysis constants of the different water-dioxane mixtures has been performed with the MINIGLASS program (Izquierdo,1986), quite similar in terms of mathematical background to SUPERQUAD and that allows the calculation of the liquid junction potential related to each solvent composition.

CHAPTER 3.

POLYNUCLEOTIDES IN WATER/DIOXANE MIXTURES.

3.1. Structure of the polynucleotides and particularities of their acid-base equilibria.

The study of the processes tied to live beings has always been one of the main goals of Science. Thus, in the ancestral Medicine, the presence of some external signs was associated with the existence of certain diseases; afterwards, the increasing knowledge of the internal anatomy helped in a great extent in the diagnostic of organic malfunctions. Towards the end of the XIXth century, pioneering work was carried out about the main constituents of alive beings. However, it was not till the XXth century that the development of new instrumental techniques facilitate the study of the molecular structure of biological constituents. This fact led to the birth of Biochemistry, a young and active research area focused on the study of the structure of biomolecules and on the comprehension of the processes in which these complex compounds are involved.

The large family of biological molecules includes an especially relevant group, the polynucleotides (Saenger,1984), (Freifelder,1987), (Adams,1992). Some of them have as fundamental missions as codifying the genetic information (DNA), transmitting this information to the ribosomes for replication (messenger RNA, mRNA) or participating in the protein synthesis by carrying amino acid residues (transfer RNA, tRNA). Several subunits of these biomolecules also play essential roles, like cyclic adenosine monophosphate (cAMP) which is the second hormonal messenger and adenosine triphosphate (ATP), which supplies energy for biological processes. The essential activity played by these natural compounds in many biochemical processes led to the use of analogous synthetic substances (modified bases, nucleosides and nucleotides) as probes in the study of biological mechanisms and as chemotherapeutical agents with proven antibiotic and antitumoural properties (Hall,1971), (Bloch,1975), (Harmon,1978), (Walker,1979), (Suhaldonik,1970;1979). The most recent tendencies have included the application of synthetic homopolynucleotides to mimic more specifically the homologous natural compounds, thus increasing the specificity in the treatment of certain diseases (Thang,1992).

The chemical behaviour of polynucleotides cannot be understood without the previous knowledge of the structure and properties of the basic constituents forming these macromolecules. A polynucleotide is a high-molecular weight biopolymer which on complete hydrolysis yields a mixture of pyrimidine and purine bases, a sugar component and phosphoric acid; partial hydrolysis gives the compounds known as nucleosides and nucleotides (see Figure 3.1.1).

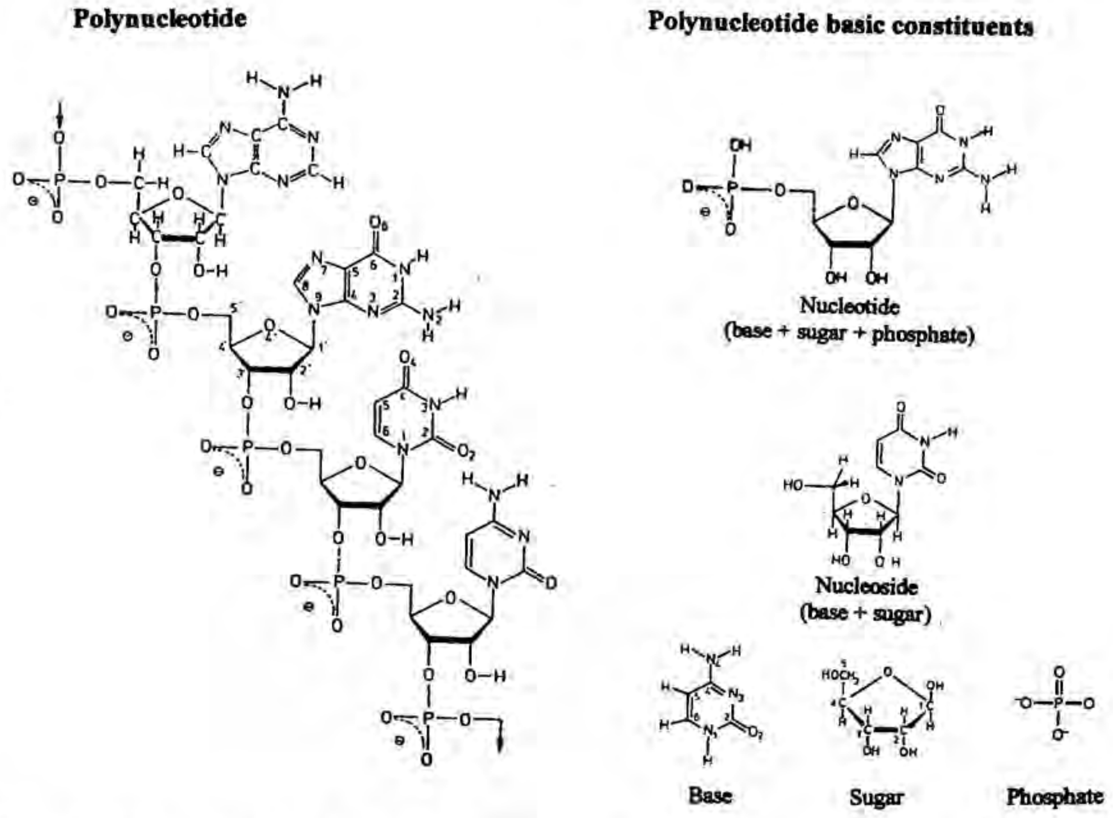


Figure 3.1.1. A polynucleotide and its basic constituents.

The monomeric unit of a polynucleotide is called nucleotide. Each nucleotide has three molecular fragments:

- **a cyclic five-carbon sugar.** This sugar can be ribose, as in RNA, or deoxyribose, as in DNA. Both sugars are in a D-furanose form, the only difference between them being the absence of the 2'-OH group in the deoxyribose. The presence of this hydroxyl group is responsible for the easier chemical and enzymatic degradation of polyribonucleotides (Adams,1992). Stereochemically, the sugar molecules adopt either twist or envelope conformations to minimize the intramolecular steric hindrance between their substituents (de Leeuw,1980), (Saenger,1984).

- **a phosphate group.** It can be attached to the 3' or 5' carbon of the sugar through an ester linkage. This group has the strong negative charge of nucleotides and presents a tetrahedral structure.

- **a nitrogen base.** It is linked to the 1' sugar carbon by an N-glycosyl bond and can adopt an *anti* or a *syn* conformation with respect to this bond (Schweizer,1971). Two main families are distinguished according to the basic heterocycle present in their structure: the purine bases and the pyrimidine bases. A common essential feature of all the nitrogen bases is their planar structure, where the amino groups are in resonance with the π system of the aromatic ring. Such a planar structure, together with a permanent dipole in the molecule caused by the presence of carbonyl and amino groups, promotes the development of hydrogen bonding and base stacking. These base-base interactions are key factors in the formation of the secondary structures of polynucleotides.

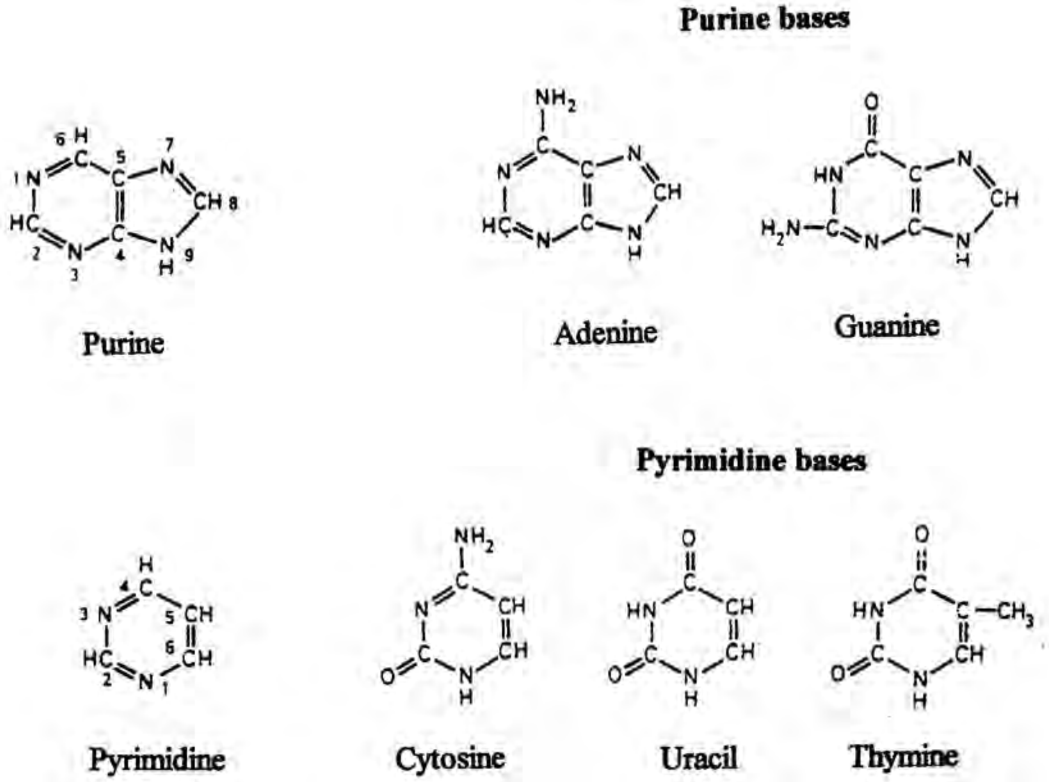


Figure 3.1.2. Some natural nitrogen bases and their basic heterocycles. a) Purine and purine bases (adenine and guanine) and b) pyrimidine and pyrimidine bases (cytosine, thymine and uracil).

The nucleotides in a polymeric chain are linked by an ester bond between the phosphate group of one of the nucleotides and the sugar 3'-OH group of the adjacent nucleotide. No covalent bonds appear between bases in the whole structure. A polynucleotide is thus formed by an alternating sugar-phosphate backbone with one phosphate terminus and one 3'-OH terminus. The partial rigidity of the sugar-phosphate bonds helps to stabilize ordered macrostructures and, at the same time, allows a certain rotation in order to reach the minimum repulsion between the constituents of the polynucleotide (Yathindra,1974), (Adams,1992).

The role of polynucleotides in many biochemical processes is conformation-dependent. This explains the usefulness of studies devoted to assessing the influence of many chemical parameters (i.e. pH, ionic strength, temperature, polarity, presence of metal ions, ...) on the structure of these compounds (Brahms,1966), (Shin,1973), (Rifkind,1976), (Causley,1982). However, most traditional studies on this subject have been carried out at fixed conditions because of the lack of suitable data treatment procedures for multivariate outputs from macromolecular evolutionary processes. One of the strongest effects on the structure of the polynucleotides stems from the introduction of hydrogen ions in solution. Therefore, the study of protonation processes from multivariate data collected during systematic variation of pH could clarify dubious results from previous studies performed at fixed pH, where the detection of coexisting species was not possible and the description of conformational transitions could only be made in a rough, qualitative and rather intuitive way.

The study of the thermodynamical and conformational transitions associated with a polynucleotide protonation process should take into account some features due to the macromolecular nature of this biomolecule (Buffle,1988) and to the properties of its basic constituents (Saenger,1984), (Burger,1990).

As a macromolecule, some side effects caused by the interaction between neighbouring protonation sites must be considered, namely:

- **polyfunctional effect.** This takes place when the monomers in the polymeric chain are different, as in DNA, and comes from the presence of diverse functional sites in the macromolecule. This functional variety causes an overlap between the several occurring equilibria and therefore, the description of the protonation process associated with each functional site in the polymer becomes more difficult.

- **polyelectrolytic effect.** This phenomenon is due to electrostatic changes in the vicinity of the protonation sites as protonation proceeds. The gradual appearance or disappearance of charges in the polymer can modify the proclivity of analogous sites to be protonated. When the polyelectrolytic effect is present, identical functional sites are no longer equivalent and the value of their equilibrium constants becomes dependent on the protonation degree of the macromolecule.

- **conformational transitions.** The spatial structure of a polyelectrolyte may change drastically because of the establishment of some new intramolecular interactions (i.e. hydrogen bonding) which becomes possible only when the sites are in a certain protonated form or simply because of spatial rearrangements oriented to minimize the electrostatic repulsions in the polymeric chain.

The chemical nature and particular stereochemistry of the basic constituents of the polynucleotides also promote certain kinds of intramolecular interaction, which are responsible for the formation and stabilization of highly ordered structures (e.g. single-, double- or multi-stranded helical conformations). There is also evidence of a cooperative mechanism between the appearance of certain conformations and the proton uptake process in some polynucleotides (Taylor,1982). A clear example is the formation of the double helix of the polycytidylic acid, polyC, in aqueous solution; thus, this helix forms only when the interstrand hydrogen bonding is established between a protonated base and a deprotonated base and, at the same time, the formation of this double helix contributes to the increasing stabilization of the protonated bases it facilitates the establishment of the hydrogen bond N-H^+····N as the helical structure grows (Kistenmacher,1978), (Casassas,1995).

The most important intramolecular interactions in polynucleotides are developed between the nitrogen bases as a result of the planar structure of these molecules. Two main groups can be distinguished:

- **hydrogen bonding**. This interaction involves pairs of coplanar bases and occurs on the same plane of the bases. The most usual hydrogen bonds are either N-H····N or N-H····O, with a donor N-H amino or imino group. The base-pairing needs at least two hydrogen bonds to close the typical cyclic structures between the two bases to be linked (Pullman,1966). It has even been suggested that such cycles could facilitate the exchange of hydrogen atoms between the donor and the acceptor centres of the associated bases. Such an atom transfer, if present, would not entail any change in the general structure of the polynucleotide. The formation of hydrogen bonds is more favourable between complementary bases (adenine:thymine and cytosine:guanine) but is also possible between identical bases.

Base-pairings are somewhat symmetrical. In double helices, if the symmetry axis is perpendicular to the bases, the sugar-phosphate backbones in each strand of the helix are equally oriented, or parallel. If the symmetry axis is parallel to the base pairs, the two sugar-phosphate backbones are antiparallel, i.e. show opposite orientations in each strand.

Polynucleotide hydrogen bonding interactions are favoured in solvents with low acidity and low basicity because of the lesser competition of the hydrogen bond donor and acceptor sites of the solvent in the formation of the intramolecular links.

- **base stacking.** This interaction is developed between bases placed in different parallel planes and is perpendicular to the base planes. It involves atomic groups, like those in the nitrogen bases, with an aromatic system and a permanent dipole caused by the presence of polar functional groups. It is more frequent in polar media, where the affinity between the organic part of the molecule and the solvent is very low. Thus, the nitrogen bases stack one above the other forming a pile that minimizes the contact surface offered to the solvent. The stacked bases show a certain spatial specificity, with the polar substituents of one base (-C-NH2, -C=N-, -C=O) placed over the aromatic system of the adjacent base. This specific orientation creates an induced dipole which strengthens this kind of union between bases (Ts'o,1963), (Bugg,1971). The mechanism of base stacking is cooperative and therefore, the energy necessary to pile up two bases decreases if these bases are already stacked with some others.

Although the stabilization of base stacking is mainly due to the dipole-induced dipole interaction caused by the polarizing effect of the permanent dipole of one base on the π electronic system of the neighbouring one, other factors, like the London dispersion forces (Pullman,1966), (Hanlon,1966) and the hydrophobic interactions are also understood to contribute in this process as well (Saenger,1984), (Freifelder,1987).

The London dispersion forces are van der Waals forces between neutral atoms. They increase with the product of the polarizability of the interacting molecules. This would explain why the pyrimidine bases stack more easily than the purine bases.

Stacking interactions are much stronger in water than in organic solvents; nevertheless, hydrophobic interactions are not always considered to be responsible for the stacking process and there is no agreement either about the mechanism of such possible interactions (Tanford,1978), (Scheraga,1978), (Hildebrand,1979), (Friedman,1995). A generalized theory holds that water, due to its low affinity for aromatic systems, is placed around the bases forming a clathrate-like structure that includes these nitrogen molecules. This configuration favours the solvent-solvent interactions more than the base-solvent ones and increases entropy. If the bases stack, the total surface exposed to the solvent decreases, the number of water molecules to be ordered around them is lower and the entropy increase is minimized.

The combined action of the hydrogen bonding and base stacking interactions explains the stability of the single and multiple helical structures of the polynucleotides (deVoe,1962), (Borer,1974), (Rodley,1976), (Saenger,1984), (Adams,1992). Thus, the tridimensional structure of the single helices is mainly supported on the base stacking interactions and on the partial rigidity of the phosphate-sugar backbone. The formation of double and multiple helices would be hardly understood without the occurrence of the interstrand unions by hydrogen bonding. The base stacking, stronger between paired bases than in single helical structures, also yields a high degree of compactness to these latter structures.

All the helical structures share a common property in aqueous solution: the hydrophilic phosphate-sugar backbone is placed in the external part of the helix, whereas the core of the helix is formed by the hydrophobic bases. This spatial arrangement limits the water access to the less polar polynucleotide constituents.

There are many possible helix symmetries. They can be left- or right-handed and the number of residues per turn and the helical pitch can be very diverse as well. A helix is regular if its symmetry axis coincides with the symmetry axis of the base pairs (e.g. double-helix of polyadenylic acid, polyA). An irregular helix presents different grooves between consecutive strands (e.g. DNA).

3.2. The hydroorganic media used as emulators of biological environments.

The large percentage of water in all living organisms justifies the generalized use of this solvent in biochemical studies. Nowadays, the type of interactions developed by water around proteins or nucleic acids is well known, although the mechanisms of these interactions are still a matter of research (Saenger,1984), (Pethig,1992). The action of water goes beyond the usual solvating function on simple solutes and has a decisive influence on the stabilization of ordered structures, such as the single- or multistranded helices of polynucleotides. These helices show a sugar-phosphate backbone at the outer edge of the macromolecule and a less polar core formed by the nitrogen bases. The electrostatic repulsion between neighbouring phosphate groups is reduced by the presence of surrounding water molecules and the low affinity of the nitrogen bases to water promotes the phenomenon of base stacking that hinders the access of water molecules around these partially aromatic molecules (Freifelder,1987), (Adams,1992).

The role of water in the polynucleotide structure is an example of the general solvent effect on these macromolecules. Variations in the solvent properties significantly affect the development and extent of the base-base interactions; thus, a decrease in solvent polarity (this term being considered as the overall solvent ability to interact with the solutes) favours the hydrogen bonding between bases due to the lower competition of solvent molecules in the establishment of these bonds and weakens the tendency of bases to stack since the affinity between the bases and the surrounding solvent increases.

There is an intrinsic interest in studying the solvent effect in the polynucleotide conformational transitions and the use of solvents with different polarities can also help to understand the causes and mechanisms of the interactions between bases. Nevertheless, these are not the only reasons to work with solvents other than water in the study of polynucleotides. Though water is by far the most suitable solvent to reproduce biological conditions, low polar biochemical microenvironments, such as active sites of enzymes, side chains in proteins located in low dielectric cavities, and others, need some other kind of solvents to be properly emulated (Sigel,1985), (Sigel,1993), (Delauzon,1994). The polarity of these microenvironments is determined analogously to the empirical solvent parameters, i.e. making the measure of some solvent-dependent solute properties in these environments and in other media. The comparison between a complexation constant measured in the active site of an enzyme and the complexation constants of a solute with the same functional groups measured in a solvent series with a wide polarity range has long been used to know indirectly the polarity of the enzyme (Sigel,1985). A more recent approach directly based on the solvatochromic parameters is the enzymichromism (Kanski,1993). This method uses solvatochromic indicators chemically similar to those proposed in the empirical scales and stereochemically adapted to be able to penetrate in the active sites of enzymes to obtain information about the microscopic properties of the enzyme.

The study of the polynucleotides behaviour in low polar solutions provides a wider vision of the role of these macromolecules in biochemical processes. There is a large number of solvents and solvent mixtures less polar than water; however, the choice of the most suitable medium for these studies should be carried out bearing in mind that the main reason to use these alternative media is the imitation of some special biological environments. Hydroorganic mixtures seem to be the ideal solvent for this purpose. On the one hand, they have a certain amount of water, a feature common to all the biological environments, and on the other hand, the presence of an organic cosolvent causes the required polarity decrease. There is also a subtle analogy between the water behaviour in the biological environments and in a polynucleotide hydroorganic solution. In biological systems, the access of water molecules to the mentioned environments is difficult. The water structure must be broken and consequently, the dipole induced water-water interactions, which usually enhance the interaction of water with solutes, are weakened. In short, the process above could be described as a "water deactivation". Analogously, in a hydroorganic solvent, the organic cosolvent performs the double function of breaking the water structure and hindering the access of water to the biomolecule.

The passive function of the cosolvent, almost limited to producing a hindrance effect on water, advises against the use of organic solvents able to create strong or specific interactions with the polynucleotides. Such solvents would give a distorted vision of the polynucleotide behaviour, completely different to what really happens in biological systems, and dependent on the specific solvent selected. Within the large group of hydroorganic mixtures, water-dioxane mixtures have been chosen because of the diversity of their characteristics and behaviour and because of the rather passive role of dioxane. Indeed, when looking at the solvatochromic parameters of both partners in the solvent mixture (see Table 1), the dioxane solute/solvent interactions are clearly weaker than those developed by water.

Table 1. Solvatochromic parameters of water and dioxane.

Solvent	α	β	π^*
Water	1.17	0.47	1.09
Dioxane	0	0.36	0.54

The only objection could be the similar basicity of both solvents. Nevertheless, the similar chemical nature of the hydrogen bonds formed by both solvents, with an oxygen atom with an sp^3 hybridation would not cause any spurious effect in the study of biological processes performed in conditions imitating biological microenvironments.

Water-dioxane mixtures have been used in biocoordination studies by Sigel et al. (Sigel, 1993), who have focused their research on the protonation and complexation of monomeric solutes involved in biochemical processes. These authors and others (Delauzon,1994) state that most biological microenvironments have dielectric constants between 30 and 70. This range of values coincides with compositions of water-dioxane mixtures between 20 and 60 % (v/v) of dioxane. This has led to the choice of the intermediate compositions of 30 and 50% of dioxane in all the experiments related to this chapter.

3.3. Selection of the polynucleotides to be studied

The study of the acid-base behaviour of some polynucleotides in water/dioxane mixtures is within the frame of a large project undertaken by the research group of Chemometrics and Solution Chemistry in the Departament de Química Analítica de la Universitat de Barcelona. This project is focused on the interpretation of the effect that various chemical factors, namely the introduction of metal ions or hydrogen ions in solution, the temperature and the solvent properties, can cause on the thermodynamical and conformational transitions of polynucleotides.

The homopolynucleotides selected have already been studied in aqueous solution. The study of the chemistry related to these homopolymers is the previous step for the future research involving natural polynucleotides with larger functional variety in their monomers, such as RNA or DNA. The biopolymers chosen are the following:

- polyurydylic acid, or polyU, where the nitrogen base in all the monomers is uracil.
- polycytidylic acid, or polyC, where the nitrogen base in all the monomers is cytosine.
- polyadenylic acid, or polyA, where the nitrogen base in all the monomers is adenine.

The diverse features of the three selected polynucleotides allow a sound assessment of the influence of the solvent properties and of the different nitrogen bases on the proton uptake processes. The interest of this work has been focused on the points listed below:

solvent effect on the macromolecular features of the polynucleotides. In aqueous solution, polyA and polyC have a clear polyelectrolytic effect, absent in the polyU protonation process. Concerning the conformational changes in aqueous solution, polyU keeps a random coil conformation during the whole protonation process, whereas deprotonated polyA and polyC are found to present a single-helical conformation which evolves to different spatial structures as the proton uptake takes place.

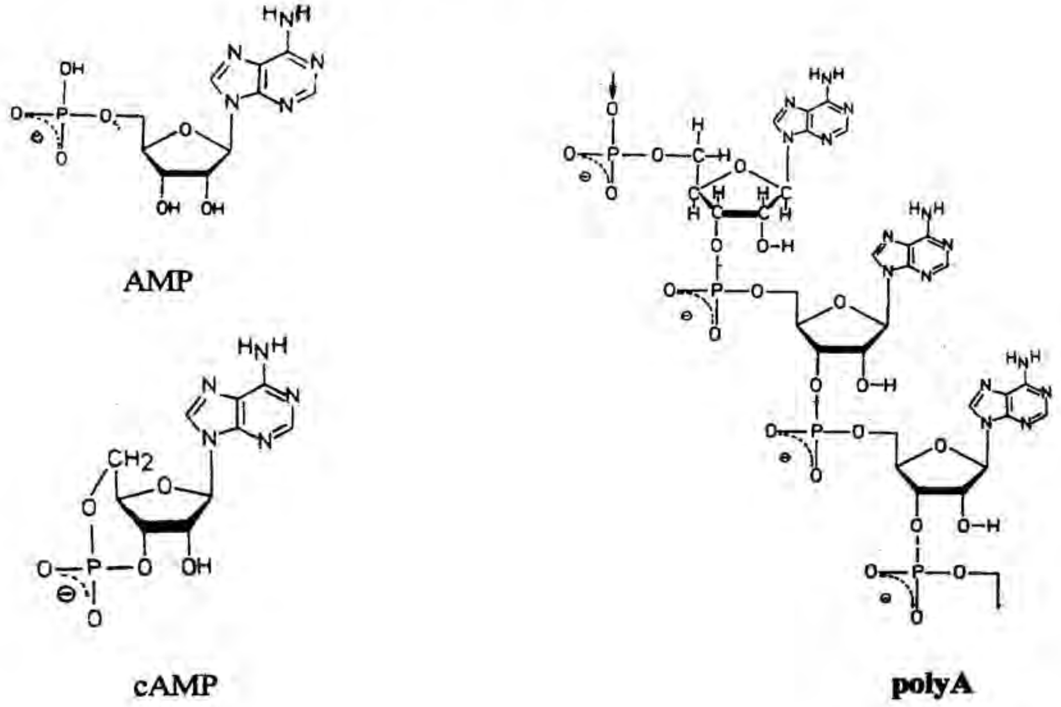
- **effect of the heterocycle present in the nitrogen base.** PolyC and polyU contain both purine bases in their polymeric chains, whereas polyA has a pyrimidine base.

- **effect of the functional group involved in the proton uptake process.** The polyU protonation site is the N(3) amide-like nitrogen present in uracil. The deprotonation of this functional group involves the appearance of a negative charge in the nucleotide. PolyA and polyC both protonate nitrogens in the aromatic rings of adenine (N(1)) and cytosine (N(3)) providing the protonated heterocycles with a positive charge.

3.4. Study of the monomeric units of the macromolecule: the model of cyclic nucleotides.

The complexity of the polynucleotide acid-base equilibria is mainly due to the macromolecular character of these biomolecules. As a general rule, the most advisable starting point in facing any study of polymeric substances consists in determining the chemistry of their monomers. To do so, model solutes similar or equal, when possible, to the monomers in the polymeric chain are used. This preliminary research provides complete information about the chemical behaviour of the functional groups and spatial structure of the monomeric units in the absence of the additional effects that can appear when these monomers are placed in a macromolecular structure (i.e. polyelectrolytic effect, conformational transitions,...). Therefore, any difference noticed between the macromolecule and the monomer behaviours can be surely attributed to the presence of the effects mentioned above.

The composition of homopolynucleotides is perfectly known; therefore, the most suitable model substance seems to be the only nucleotide present in the macromolecule. However, the structure of an isolated nucleotide differs from the structure of this molecule in a polynucleotide chain. The phosphate group of the isolated nucleotide is linked to a ribose ring, whereas the phosphate group in the polymeric chain bonds the phosphate group to its own ribose ring and to the ring of the adjacent nucleotide. As a result, the isolated nucleotide shows one additional free oxygen that can also be protonated. Such a difference in structure is not noticed when cyclic nucleotides are used as model substances because they show the same functional groups and also the same protonation sites as the polynucleotide monomers, as shown in Figure 3.4.1.


AMP

cAMP

polyA

Figure 3.4.1. Structure of the nucleotide adenosine monophosphate isolated (AMP), cyclic (cAMP) and in the polyA chain

The free oxygen atom in the isolated nucleotide is blocked in the cyclic monomer owing to the formation of a closed cycle in which this atom is linked to the 3' position of the ribose ring. Furthermore, the more rigid conformation of the sugar-phosphate bond in the cyclic nucleotide caused by the presence of the intramolecular cycle is more similar to the situation of the sugar-phosphate backbone in the polynucleotide chain.

The experimental study of the acid-base equilibria of cyclic nucleotides has been carried out potentiometrically using the experimental setup and data treatment for small solutes in hydroorganic media recommended in chapter 2.

3.5. Study of the homopolynucleotide

In all the work presented, the acid-base equilibrium of a polynucleotide has been interpreted as the equilibrium process of the protonation sites in its monomeric unit. Therefore, the reaction taken into account in a protonation process is:

$$\text{deprotonated monomer} + H^{+} \leftrightarrow \text{protonated monomer}$$

and the related protonation constant is

$$K = \frac{[\text{protonated monomer}]}{[\text{deprotonated monomer}][H^{+}]}$$

The usual concepts in the acid-base equilibria related to the acid concentration also change accordingly. Thus, the protonation degree is now defined as:

$$\alpha_p = \frac{[\text{protonated monomer}]}{[\text{total monomer}]}$$

and the equation to calculate the protonation constant retains the same form as for simple solutes:

$$K = \frac{\alpha_p}{\left(1-\alpha_p\right)\left[H^+\right]}$$

The definition adopted for the acid-base equilibria of polynucleotides leads to the expression of the polynucleotide concentration in moles of nucleotide per volume unit. In the absence of macromolecular effects and only in this absence, the numerical treatment of the data coming from potentiometric or spectrometric titrations can be carried out by using traditional least-squares curve fitting methods based on the previous postulation of a fixed chemical model.

The acid-base equilibrium of a macromolecule is as simple as the same process for a small solute when the protonation sites in the polymeric chain are equivalent, i.e. when the polyelectrolytic effect is not present. However, this is not the most common situation and the non-fulfilment of the mass action law is very frequent in macromolecular structures. Indeed, finding two identical protonation sites is difficult once the proton transfer process has started. The evolution of an acid-base equilibrium always implies the appearance or disappearance of charges in the polymer and this fact causes variations in the electric field on the surface of the molecule that affect the later development of the reaction. It is understandable that, due to the proximity of analogous sites, the tendency of a neutral site to be protonated with the two neighbouring sites deprotonated may not be the same as that of one site affected by the repulsion of two protonated neighbouring sites. Other factors that can also alter the geometry and intensity of the electric field in the macromolecule are hydrogen-bonding interactions and changes in the spatial structure of the polynucleotide.

The calculation of a protonation constant in each successive stage of the reaction is used to identify a polyelectrolytic effect. If the log K values obtained along the whole protonation process are equal to each other, there is no polyelectrolytic effect and the acid-base equilibrium can be defined with a sole thermodynamical constant, as is done with small molecules. Changes in the log K values with the extent of the reaction arise when the protonation sites are not equivalent, i.e. when the polyelectrolytic effect is present. The K values related to these latter processes lack thermodynamical meaning and are called apparent constants (K_{app}). Each of these apparent constants should be considered as a macroscopic average of the microscopic thermodynamical constants associated with each of the sites reacting between two consecutive stages of the reaction.

In a protonation process with polyelectrolytic effect, the variations of log K_{app} are related to the changes in the protonation degree, α_p:

$$\log K_{app} = f(\alpha_p)$$

where $f(\alpha_p)$ can be a linear or a non-linear function, depending on the pattern of the polyelectrolytic effect. The acid-base behaviour of the polymer is usually characterized by the constant calculated through the extrapolation of the mathematical function $f(\alpha_p)$ for $\alpha_p =$ 0. This theoretical value, which is the intercept of the mathematical model obtained, is called intrinsic constant (K_{intr}) and represents the hypothetical value of the equilibrium constant associated with the protonation site in the polymeric chain when there are no macromolecular effects caused by neighbouring protonated sites.

3.5.1. Potentiometric monitoring.

Potentiometry is the most accurate technique to study equilibrium processes. The suitable treatment of the e.m.f. readings obtained in the acid-base titration of a polynucleotide provides information about the transitions between the different chemical species involved in the protonation process, namely:

a) the distribution plot of the chemical species, where the evolution of the concentration of the differently protonated species with the pH is shown,

b) the value of the protonation constant in each titration point, and

c) the detection of the presence and pattern of the polyelectrolytic effect, through the observation of the evolution of the protonation constants with the protonation degree.

Though potentiometry does not provide information about the spatial structure of the chemical species involved in the protonation process, the detection of phenomena like the polyelectrolytic effect is very often an indication of the occurrence of conformational transitions linked to the protonation uptake process. In this way, potentiometry also furnishes hints related to the structural aspects of the acid-base equilibrium.

The main drawback associated with the potentiometric technique is the relatively high concentration levels of solute required to give reliable results. This has prevented the application of this technique in the study of polynucleotides with a low solubility in water/dioxane solutions.

All the quantitative information mentioned above in relation with the polynucleotide protonation processes is obtained by performing the suitable calculations with each e.m.f. reading of the potentiometric titrations separately. In the absence of polyelectrolytic effect, the least-squares curve fitting method explained in chapter 2 is applied afterwards to obtain the thermodynamical constant associated with the protonation process of the polynucleotide.

3.5.2. Spectrometric monitoring

The polynucleotide activity in many biochemical processes is clearly conformation-dependent. Base-base interactions, such as hydrogen bonding and base stacking, are mainly responsible for the variety and singularity of the polynucleotide conformation; therefore, chemical parameters causing modifications in these interactions also promote variations in the polynucleotide activity. The interactions between bases are strongly pH-dependent; thus, base sites involved in a protonation process change noticeably the tendency to form hydrogen bonds depending on their protonation state and experimental data have confirmed that the increase in ionic charge produced by protonation or deprotonation processes induce a stacking decrease between pyrimidinic bases (Saenger,1984).

There is a clear connection between the acid-base equilibria and the conformational transitions of a polynucleotide: sometimes a certain protonation state favours a given conformation; at other times, a certain conformation modifies the reaction tendency of a protonation site. This is why unusual variations in an acid-base equilibrium (i.e. presence of polyelectrolytic effect) can become an indirect tool to detect possible conformational transitions.

Nevertheless, the description of the pH-dependent structural changes of a polynucleotide goes beyond the detection of possible conformational transitions and is focused on the complete differentiation and identification of all the stereochemical species present before, during and after the protonation processes. Despite the often close relation between protonation states and conformational transitions, the assumption of a general one-to-one correspondence between chemical species and conformers is not warranted; i.e. the number of conformers can be different from the number of chemical species. Indeed, two different chemical species can share the same conformation and, on the contrary, a sole chemical species can present more than one conformational form.

The simultaneous acquisition of e.m.f. readings, related to the protonation changes, and spectra, related to the structural changes of the macromolecule, is the best option to avoid incorrect correspondences between conformers and chemical species. Such a working procedure is the base of the spectrometric titrations, whose experimental setup is presented in Figure 3.5.2. In these experiments, the spectrum and the pH value of the polynucleotide solution are measured after each titrant addition (acid or base, according to the case) until the desired pH range is covered.

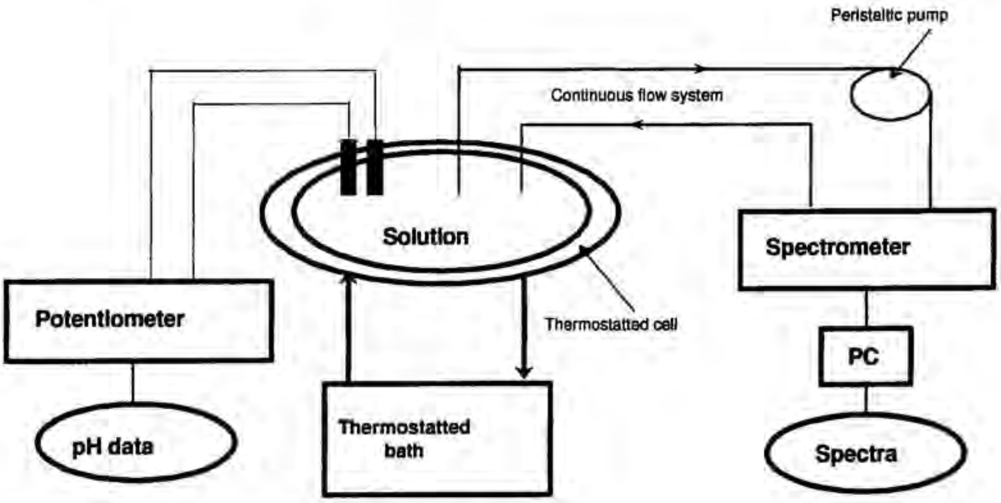


Figure 3.5.2. Experimental setup of a spectrometric titration.

The information obtained after treating all the spectra collected in a spectrometric titration includes:

- the concentration of each absorbing species at each titration point (i.e. at each pH value).
- the pure spectrum of each absorbing species.

The distribution plot (concentration vs. pH) of all the absorbing species is used to identify those involved in the proton uptake and those that are related to a conformational transition without any associated chemical reaction. The structural changes linked to protonation usually take place usually in narrower and more reproducible pH ranges than the conformational changes due only to spatial rearrangements of the macromolecule. The results obtained from potentiometric titrations and the information related to the studies of their cyclic nucleotide also indicate which absorbing species are related to each protonated state of the macromolecule.

Once the correspondence among conformers and chemical species is established, the protonation constant in each point of the titration can be calculated as:

$$K = \frac{\sum[\mathit{protonated\ conformers}]}{(\sum[\mathit{deprotonated\ conformers}])[H^+]}$$

Several techniques are available for the study of the structure of polynucleotides. Most experiments to determine the crystalline structure of these macromolecules are carried out with solid phase X-ray diffraction techniques. Proton nuclear magnetic resonance and infrared spectrometry provide information about polynucleotides in solution. Both techniques, however, require deuterated solvents: in the case of the ^{1}H NMR, this is an intrinsic requirement of the technique, whereas in the IR spectrometry the use of deuterated solvents is necessary due to the overlap between the spectral absorption ranges of the polynucleotides and of the water molecule (H_2O). Spectrofluorimetry has also been used to confirm the presence of certain conformations, but it cannot be generally applied to detect all the species in solution.

UV/Visible spectrometry and circular dichroism spectrometry (CD) were chosen for the structural study of these substances in solution. These techniques share the following features:

- high sensitivity to variations in the base-base interactions and to the electronic transitions associated with the acid-base equilibria,
- detection of all the polynucleotide conformations,
- ability to record the spectra in the usual working solvent, and
- instrumental responses separable in linear additive terms, each of them related to a different absorbing species.

There is a close relation between the information given by both spectrometries. Thus, the zones of spectral absorption of the polynucleotides in both techniques coincide and, when conformational transitions are monitored, the appearance of an optical phenomenon in one of them is always linked to the simultaneous emergence of a different optical phenomenon in the other one. The interrelation between the optical activity in both UV/Vis and CD spectrometries can be explained because in the UV absorption region, a chromophore (the nitrogen base) substituted by a chiral moiety (the ribose ring) displays a Cotton effect. Consequently, any structural transition of a polynucleotide can be noticed through the synchronized variation of the UV and the CD signals. This double detection helps to confirm any conformational change in the macromolecule.

3.5.2.1. UV-visible spectrometry

Acid-base equilibria is often monitored using UV/Vis spectrometry. Differently protonated species of a single molecule have different electronic transitions and therefore, different absorption spectra. A protonation process can be followed spectrometrically thanks to the appearance of a consistent pH-dependent blue or red shift in the recorded absorption band of the molecule as the evolution between the chemical species involved in the acid/base equilibrium takes place.

UV/Vis spectrometry is applied to study the polynucleotide acid-base equilibria and to detect their related base stacking variations. The intensity and extent of this base-base interaction determine the appearance of more or less ordered polynucleotide structures; therefore, the observation of the base stacking changes occurring during any conformational transition is enough to describe this structural process. Two stacked bases couple their chromophores and the intensity of their absorption band is lower than the intensity that the same spectral band would present if they were isolated molecules. Thus, a conformational transition leading to a more ordered structure (i.e. with a more intense base stacking, e.g. single helix → double helix) can be detected through the appearance of a hypochromic effect. In contrast, hyperchromism is caused by the disappearance of the interactions established between the π electronic systems of adjacent bases, a direct consequence of the base unstacking linked to any disorder in the polynucleotide structure (e.g. a helix → random coil transition). Any kind of polynucleotide conformational transition, irrespective of the chemical variable responsible for it (pH, solvent, T, ionic strength,...), can be monitored through the changes in the absorption intensity of the nitrogen base spectral band, placed in the wavelength range around 260 nm.

The optical phenomena related to a polynucleotide acid-base equilibrium, though appearing in the same spectral zone, do not interfere with each other. Whereas the proton transfer shifts the spectral band, the pH-dependent conformational transitions only affect the band intensity. Both processes can then be simultaneously followed owing to their different related spectral manifestations.

3.5.2.2. Circular dichroism

Many substances rotate the plane of a polarized light. These compounds all present molecular or crystalline asymmetry and are said to have optical rotatory power. If an asymmetric molecule can also absorb the two vectorial components of a polarized light beam differently, the substance presents circular dichroism. Measuring the rotation of the polarization plane as a function of wavelength is the basis of the optical rotatory dispersion technique (ORD), whereas combining the measure of the rotation and the differential absorption of the two components in the polarized light beam is the information used in circular dichroism spectrometry (CD).

The ellipticity angle, θ, is the experimental variable related to CD spectrometry. This angle appears as a consequence of the difference arising between the absorptivities associated with the two components of the polarized beam and is expressed as follows:

$$\theta = \frac{\pi\left(\varepsilon_R - \varepsilon_L\right)}{\lambda}$$

where ε_R and ε_L are the absorptivities of the two polarized components and λ is the working wavelength. θ is an absorption measurement, though differential in nature, which can be described by the following Beer-Lambert-like expression:

$$\theta = [\theta]\ l\ c$$

which linearly relates the ellipticity to the concentration. [θ] is the molar ellipticity and l is the optical pathlength. ε_R, ε_L and consequently, θ, depend on the working wavelength. The plot showing the variation of θ vs. λ is a CD curve or spectrum.

ORD and CD detect the presence of chiral centres in a molecule and their experimental curves are also related due to the similarity of the basic optical phenomena controlling both techniques. ORD gives two types of signal: those coming from the normal light scattering, the so-called *soft curves*, and those due to the anomalous light scattering, or *Cotton effect curves*. CD spectrometry only detects the signals due to Cotton effect because these are the only ones involving spectral absorption, one of the essential requirements of circular dichroism. Concerning the quality of the structural information, the Cotton effect curves yield more useful responses. The specificity of this signal for the different chromophores, higher in CD than in ORD, has led to the choice of the former technique to carry out all the chiroptical studies presented in this project. Priority to the ORD application would only be given if the molecule to be analyzed did not have any absorbing centre. In this case, the molecule would not yield any CD signal, whereas some information could be extracted from the ORD soft curves.

CD has been used successfully in the study of many thermodynamic and conformational studies of optically active substances. Despite its wide application, the theoretical basis of circular dichroism is not very clear. Complex molecular theories and empirical rules valid for limited families of substances are being developed, but the complete understanding of the optical activity of an unknown compound is not yet possible. The most usual procedure to interpret CD spectra is comparing the spectrum of the analyte with those of similar substances with known configurations. The lack of sound rules to explain the CD data advises the use of additional structural techniques to confirm the configuration of the substances analyzed.

For polynucleotides, the circular dichroism signal comes from the presence of a chromophore (nitrogen base) with a chiral substituent (ribose ring) in each of the monomeric units of the biomolecule. The glycosyl bond is the rotation axis linking the base and the sugar and the different orientations of these molecular fragments around this axis are responsible for the main changes detected in the CD spectra.

The polynucleotide CD signals appear in the same spectral region of the UV/Vis signals because there is no possible circular dichroism in wavelength ranges at which the molecule does not absorb. The position of the spectral bands in both spectrometries coincides although the shapes are not always similar. This means that any polynucleotide process usually followed through the UV spectral band shifts can be equally analyzed looking at the shifts in the analogous CD bands. This is the case of the acid/base equilibria, where the protonation/deprotonation of the macromolecule is associated with a red or blue shift in the absorption band.

The pH-dependent conformational transitions of polynucleotides can also be detected with the CD spectrometry. Whereas the variations in base stacking are detected in UV spectrometry, the changes in the orientation of the ribose ring and the nitrogen base are the indicators of structural changes in CD spectrometry. The optical phenomena associated with the conformational transitions are opposite in the two spectrometries. Thus, transitions leading to more disordered structures cause the appearance of a hyperchromic effect in UV and a hypochromic effect in CD. When the evolution of the transition involves an ordination of the macromolecular structure, the optical CD and UV phenomena both change accordingly.

3.5.3. Data treatment

The spectrometric data sets related to equilibrium processes of small solutes have been successfully interpreted with the application of least-squares curve fitting methods based on the previous postulation of a chemical model (Legget,1977;1983), (Henry,1997). The well-known chemistry of these simple molecules allows the input of models where the stoichiometry of all the monitored reactions and an estimate value of their related equilibrium constants can be defined. An iterative refinement of the initial chemical parameters is carried out by using a suitable least-squares algorithm (i.e., non-negative least squares for UV spectral data (Lawson,1974) or classical least-squares for spectral data with experimental negative values, like CD) to minimize the function:

$$U = \Sigma\, \Sigma\, (A_{ij(exp)} - A_{ij(calc)})^2$$

where $A_{ij(exp)}$ is the experimental absorbance recorded at the i^{th} pH value and the j^{th} wavelength and $A_{ij(calc)}$ is the absorbance calculated for the same experimental conditions by using the model being refined. When the iterative process has finished, the thermodynamic constants related to all the equilibrium processes and the unit spectra of all the chemical species involved are obtained. All these classical model-based procedures inherently assume:

- **the fulfilment of the mass action law** (i.e., the validity of a fixed equilibrium constant during the whole reaction) and,
- **the one-to-one correspondence between absorbing species and chemical species.**

These two assumptions are responsible for the general failure of these methods when dealing with spectrometric data connected with macromolecular equilibria (e.g., those from polynucleotides, proteins,...). The greater complexity of these polymeric solutes and their related processes leads to consider the possible presence of the side effects previously referred breaking the normal rules of simple equilibria, namely:

- **polyelectrolytic effects.** The mass action law is no longer valid if the tendency of analogous sites to react in the course of the reaction is somehow modified. If this happens, log K = f(α_p), where α_p denotes the extent of the reaction. The input of mathematical expressions to model this function is not yet possible because no information about the presence and the pattern (i.e., linear or non-linear) of the polyelectrolytic effect is available in advance for an unknown macromolecular process.

- **conformational transitions**. Macromolecular biomolecules can show conformational transitions associated with a chemical reaction or with a spatial rearrangement of the molecule. If the latter phenomenon takes place, the number of structural species exceeds the number of chemical species and the assumption of a one-to-one correspondence between chemical species and spatial configurations is illicit.

The unfeasibility of proposing a sound chemical model able to describe all the thermodynamic and conformational phenomena linked to a macromolecular equilibrium excludes the use of hard-modelling methods to handle these spectrometric data sets. In contrast, soft-modelling methods are suitable to work with data sets for which neither behaviour models nor prior information about the number and identity of the species are available. Some of these procedures explore the data set and give information on the total number of species present in the data set (e.g., principal component analysis, PCA), the regions of existence of these species (e.g., evolving factor analysis, EFA) or the number of species present in each region of the data set (e.g., fixed-sized moving window evolving factor analysis, FSMW-EFA); some others, the so-called curve resolution methods are used to identify the contributions related to each pure species (e.g. concentration profile and spectrum) starting from the mixed information in the original data set. The only requirement for the use of all these mentioned soft-modelling methods is that the contribution from each pure instrumental response to the overall measurement behaves following an additive model or, in other words, that the data matrix to be analyzed must be bilinear. The combined use of exploratory and resolution methods is the strategy proposed in this project to study the spectrometric data sets related to the acid-base polynucleotide equilibria.

The experimental data collected during an acid-base spectrometric titration can be organized in a data matrix **D**, whose rows are the spectra recorded at each pH value. The spectra in the data matrix are sorted as they were measured during the titration.

$$\begin{array}{c|ccccc|}
 & \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_n \\
\hline
pH_1 & A_{11} & A_{12} & & \cdots & A_{1n} \\
pH_2 & A_{21} & A_{22} & & \cdots & A_{2n} \\
 & & & & & \\
pH_m & A_{m1} & A_{m2} & & & A_{mn} \\
\hline
\end{array}$$

Figure 3.5.3.1. Data matrix from an acid-base titration.

Owing to the evolutionary character of the equilibrium process and to the nature of the instrumental response, the matrices coming from the acid-base spectrometric titrations have the special features listed below:

- **presence of structure along the rows and along the columns.** The rows and columns of the data matrix are not randomly ordered. Whereas the rows account for the spectral information, the columns show profiles related to the evolution of the concentration of the species involved in the equilibrium process. The sequential appearance and disappearance of the chemical species involved in these processes allows the use of special techniques that take advantage of this evolutionary tendency, such as EFA and FSMW-EFA.

- **the overall variation of the data matrix can be described by a linear model of additive contributions.** The fulfilment of the Beer-Lambert law allows the expression of the total absorbance data matrix **D** as the sum of the *n* absorbance matrices corresponding to the pure spectral contributions of the *n* absorbing species present.

$$\mathbf{D} = \mathbf{A}_1 + \mathbf{A}_2 + \mathbf{A}_3 + \ldots + \mathbf{A}_n$$

Each of these *n* data matrices is the result of the outer product of two vectors, $\mathbf{c}_i$ and $\mathbf{s}_i^T$, where $\mathbf{c}_i$ is the concentration profile of the ith species (i.e., the vector having the concentration values of the ith species at each pH value) and $\mathbf{s}_i^T$ is the unit spectrum of the ith species (i.e., the vector having the values of the molar absorptivities of the ith species at each wavelength scanned).

$$\mathbf{A} = \mathbf{c}_1\,\mathbf{s}_1^T + \mathbf{c}_2\,\mathbf{s}_2^T + \mathbf{c}_3\,\mathbf{s}_3^T + \ldots\,\mathbf{c}_n\,\mathbf{s}_n^T$$

This expression can be written in a more compressed way by column-appending the *n* $\mathbf{c}_i$ vectors and row-appending the *n* $\mathbf{s}_i^T$ vectors, giving:

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}}$$

where **C** is a matrix whose columns are the *n* concentration profiles and $\mathbf{S}^{\mathrm{T}}$ is a matrix whose rows are the *n* unit spectra. This last equation is the basis of all the curve resolution methods, which can be applied only when the decomposition of the data matrix makes sense chemically, i.e., when the instrumental response of one species needs only one additive term to be described.

3.5.3.1. Exploratory analysis of a spectrometric data set.

The exploratory analysis of a data matrix provides complementary information, based on the internal structure of the data set to the chemical knowledge of the researcher about the experiment performed. The many methods devoted to finding out the hidden information in the data matrix can use either the purest real variables in the data matrix (i.e., simple-to-use interactive self-modelling mixture analysis, SIMPLISMA (Windig,1991) or orthogonal projection approach, OPA (Cuesta-Sánchez,1996)) or abstract variables, like the three methods reported below.

3.5.3.1.1. Principal Component Analysis (PCA).

This method describes the real multidimensional data space by means of some abstract variables known as principal components, PC, which are linear combinations of the original variables in the real data space. PCA leads to the compression and decorrelation of the original data space. Thus, each PC is linearly independent from the others and only a few are necessary to describe the total variation of the data set (Malinowski,1991).

In an acid-base spectrometric titration, the independent contributions to the variation of the data set come from the different absorbing species involved in the equilibrium monitored; therefore, the number of significant PCs could apparently be associated with the total number of species present in solution. Nevertheless, in any instrumental monitoring of a process other phenomena may also be responsible, though to a lesser extent, for the total variation of the experimental data, e.g. instrumental drift, baseline contributions or heteroscedastic noise. Rank analysis methods like those described in section 1.4.2. assume that the matrices are formed by experimental measurements which possess normally distributed error and tend to overestimate the number of species in the data set if the identification between number of chemical species and number of significant PCs is accepted (ref. rank analysis). No rank analysis procedure can distinguish clearly between chemical and non-chemical contributions to the total variation in a data set. Generally speaking, the more rigid the criterion (statistical or empirical) to determine the number of PCs, the greater the number of non-chemical contributions considered significant to the data set. The most recommendable strategy to determine the number of chemical species in an acid-base titration would be the use of graphical methods based on the visual comparison of the eigenvalues related to the different PCs. The magnitude of each eigenvalue is related to the importance of the source of variation represented by it. As the variation due to chemical contributions is usually much higher than the variation of non-chemical contributions, the presence of a cut-off in the plot showing the magnitudes of the log(eigenvalues) vs. the PC number indicates the separation between the PCs related to chemical contributions and those associated with other sources of variation. Even in this last case, the number of species obtained should not be considered a fixed parameter, but a probable value that can be modified if the results obtained with the later application of curve resolution methods lack chemical sense.

3.5.3.1.2. Evolving Factor Analysis (EFA).

This factor analysis-based method was first proposed by Maeder et al. to treat spectrometrically monitored equilibria (Maeder,1985;1987) and was afterwards applied to any other kind of data set coming from dynamic chemical processes (Keller,1991), such as HPLC-DAD, kinetic transformations, structural transitions,...

The EFA method plots the eigenvalues obtained from the PCA of gradually growing submatrices of the original data set. PCA is performed repeatedly in submatrices generated by increasing the size of the previous submatrix by one row . This size increase is performed from the top to the bottom of the original data matrix **D** (forward EFA) and also in the opposite sense (backward EFA). Thus, forward EFA first carries out PCA on rows 1 and 2 of the data matrix, then on rows 1,2, and 3 and so on until the final PCA on the n rows of the whole data matrix **D.** Backward EFA starts by performing PCA on the rows n and (n-1), then adds the row (n-2) to the analysis and so forth until the final PCA with the whole data matrix.

The EFA plot includes the evolution of the eigenvalues (EV) as the submatrix analyzed grows, i.e., as the evolutionary process in the data matrix goes from the beginning to the end (forward EFA) or vice versa (backward EFA). The log(eigenvalues) are plotted vs. the variable responsible for the evolutionary process (e.g. pH in an acid-base equilibrium or retention time in an HPLC-DAD data set) as shown in figure 3.5.3.1.2.1.

An EFA plot gives the following information:

- **the total number of absorbing species in the data set.** The eigenvalues related to absorbing species (or more generally, to species with a pure instrumental response) are higher than the eigenvalues attributable to noise. At the bottom of the plot, the many overlapping noise eigenvalues define a graphical threshold that can be considered the noise level of the system. All the eigenvalues above this limit are chemically significant. The total number of absorbing species can be determined from the number of significant eigenvalues obtained in the last PCA performed in both forward and backward data analyses (i.e., with the results in the right and left extremes of the forward and backward EFA plot, respectively).
- **the zones of appearance and disappearance of the absorbing species.** The appearance of a new species is detected in the forward EFA through the emergence of a new significant eigenvalue. Thus, when the row added to a certain submatrix includes a new species, the independent information provided by this species to the data set analyzed will cause a significant increase in the until then first noise eigenvalue. The chemical variable (e.g. pH, time,...)

corresponding to the new row included in the analysis indicates the chemical conditions in which the new species appears. When performing backward EFA, the process is followed from the end to the beginning. Therefore, the presence of a new significant eigenvalue indicates the row, and consequently the chemical conditions, in which a species disappears.

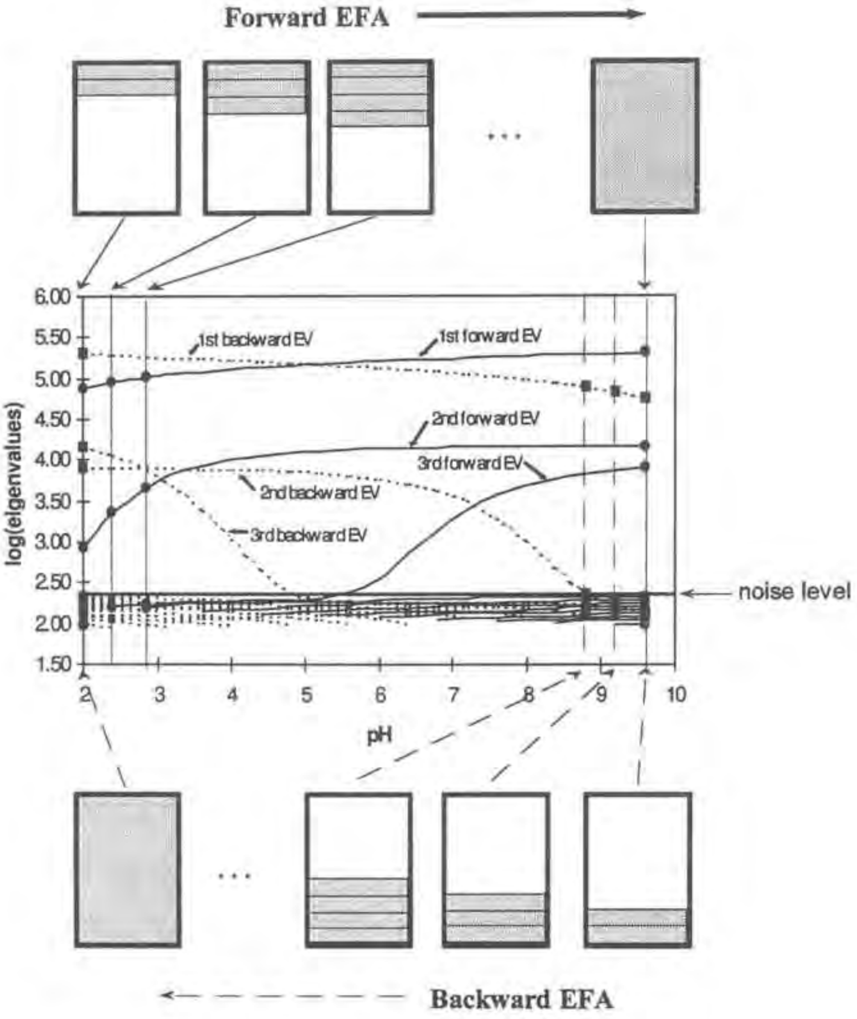


Figure 3.5.3.1.2.1. Construction of an EFA plot related to the spectrometric titration of a diprotic acid (simulated data). The continuous lines join the EV (circles) obtained in each PCA of forward EFA and the dashed lines connect the EV (squares) obtained in each PCA of backward EFA.

The evolution of the eigenvalues in the forward and backward EFA is connected to the formation and decay curves of the absorbing species in solution, respectively. Taking into account the sequential appearance and disappearance of the differently protonated species in acid-base equilibrium, the suitable connections between the lines corresponding to significant eigenvalues appearing (in forward EFA) and disappearing (in backward EFA) can provide additional information, such as:

- **an abstract distribution plot of the species involved in the acid-base equilibria.** In any acid-base equilibrium, the transitions between differently protonated species follow a sequential and logical order as the pH values increase or decrease. Therefore, an abstract distribution plot can be obtained by connecting the first species appearing in the EFA plot (i.e., the first eigenvalue arising from the noise level in the forward EFA) with the first species disappearing (i.e., the last eigenvalue appearing in the backward EFA) and, in general, the ith species appearing with the ith species disappearing until the lines corresponding to all the species present are properly connected, as shown in Figure 3.5.3.1.2.2. These abstract concentration profiles are often used as initial estimates in iterative resolution methods.

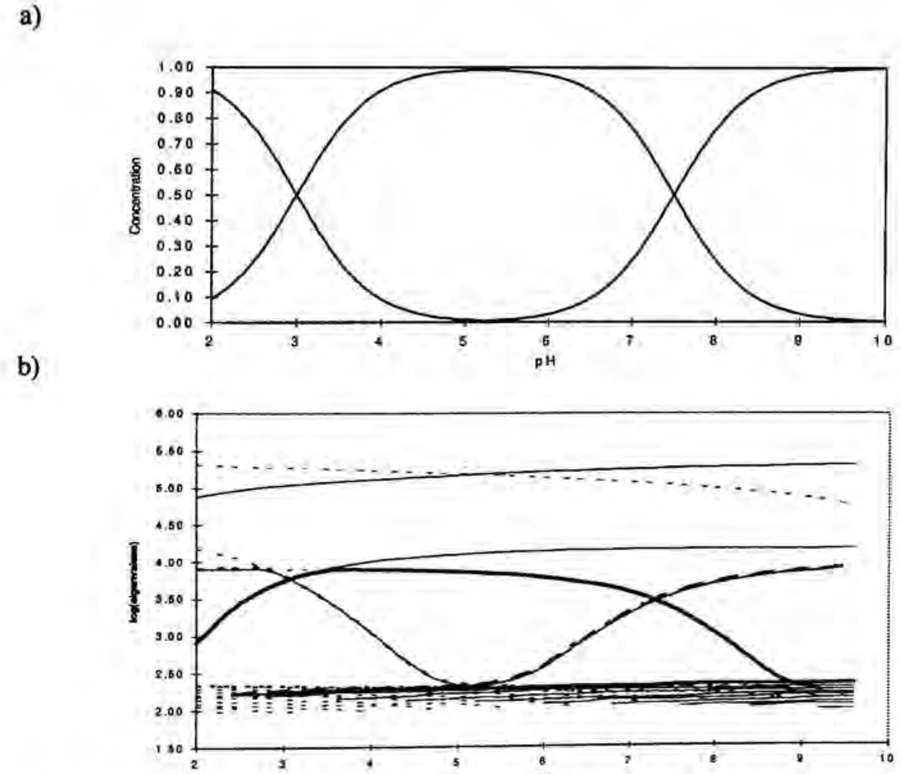


Figure 3.5.3.1.2.2. a) Real distribution plot and b) EFA distribution plot for a diprotic acid (simulated data used in figure 3.5.3.1.2.1).

- **the zones of existence of the different species.** These regions of existence (generally known as concentration windows) become evident once the distribution plot has been traced and then indicate the pH ranges at which the different species form. A very important information that can be extracted from this plot is the presence and location of selective zones (i.e., regions of the data matrix where only one species is present). The existence of these regions is essential in the later resolution of the data set.

The EFA method can also be applied in the spectral direction of the data matrix (i.e., working with the columns of the original data matrix). However, due to the usually large overlap between the spectra of the different species in solution, less information is provided, and it refers only to the total number of species in the data matrix and to the presence of spectral selective zones.

3.5.3.1.3. Fixed-Size Moving Window-Evolving Factor Analysis (FSMW-EFA).

This method was proposed by Keller and Massart as a derivation of EFA (Keller,1991) and, like this method, is especially suitable for the study of data matrices including evolutionary processes. FSMW-EFA scans the original data matrix from top to bottom by performing PCA on equally sized submatrices. A new submatrix is built by removing the first row of the previous submatrix and adding the following row in the original data matrix, i.e., moving a window of a fixed number of rows one row downwards. The size of the window usually exceeds the number of species by one, though more information can be obtained if FSMW-EFA is applied several times using different window sizes (Toft,1993).

A local rank map (i.e., a graph showing the number of coexisting species in each zone of the data matrix) is obtained by plotting the eigenvalues obtained in each PCA vs. the variable responsible for the evolution of the process, as shown in Figure 3.5.3.1.3.1. In this plot, the noise level is also defined graphically by the zone in which the noise eigenvalues appear together. The emergence of one significant eigenvalue from the noise level zone indicates the presence of only one species in that zone; the presence of two significant eigenvalues defines the zone where two species coexist, three significant eigenvalues would detect the overlap of three species, and so on.

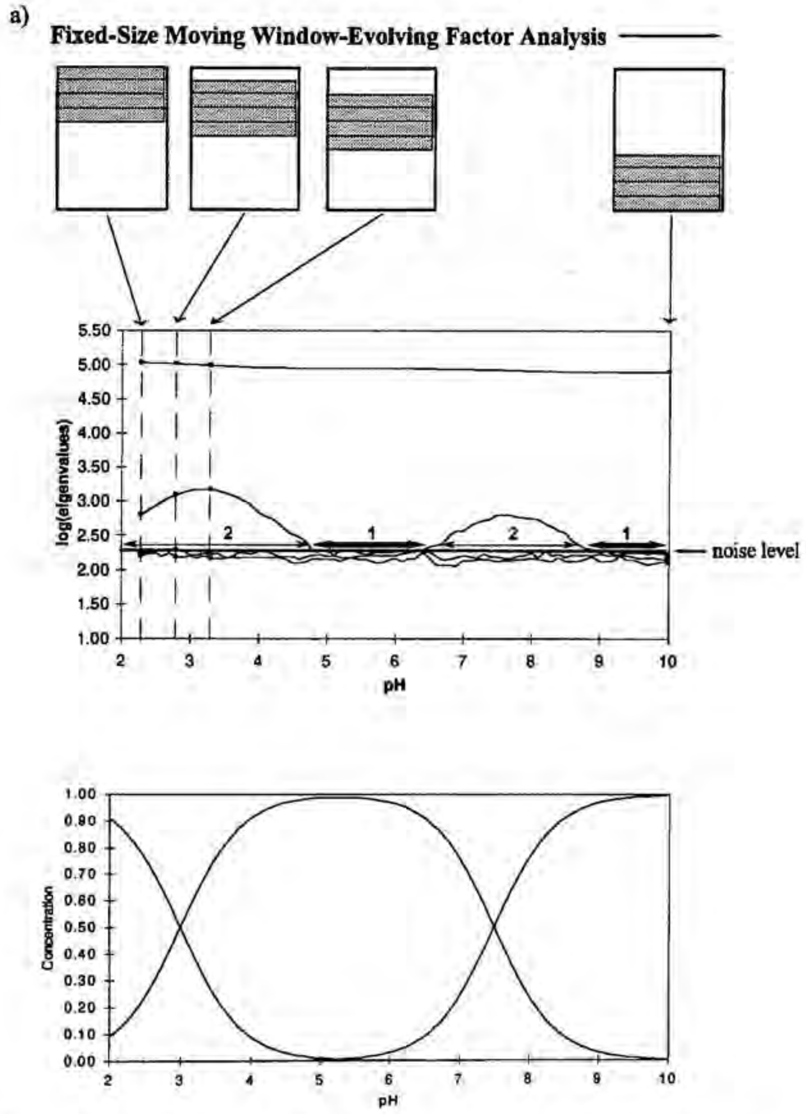


Figure 3.5.3.1.3.1. a) FMSW-EFA plot related to the spectrometric titration of a diprotic acid. The figures in the diagram indicate the number of species in the zones defined by the lines. The thick lines mark the selective zones. Circles are the EV obtained in the PCA of each submatrix. b) Real distribution plot of the diprotic acid.

The local rank map obtained through the application of FSMW-EFA provides the following information:

- **detection of selective zones.** The zones where only one significant eigenvalue is present are the selective zones of the species. The detection and location of these regions helps in the later resolution of the data matrix. Spectral selective zones can also be detected if FSMW-EFA is performed in this matrix direction.

- **detection of minor constituents.** FSMW-EFA was initially proposed to deal with peak purity problems in chromatography. The local analysis of the data matrix and the clear graphical information contribute to the detection of minor species, although they are embedded under major constituents.

3.5.3.2. Resolution of a spectrometric data set.

The multivariate output of a spectrometric titration forms a double-structured bilinear matrix containing mixed information on the evolution of all the absorbing species present in successive stages of the monitored reaction. Curve resolution (CR) methods are chemometrical procedures for the treatment of bilinear multicomponent data matrices. The ultimate goal of the curve resolution methods is the decomposition of the initial mixture data matrix **D** in the product of two data matrices **C** and **S,** each of them including the pure response profiles of the *n* mixture components associated with one of the directions of the initial data matrix (see Figure 3.5.3.2.1).

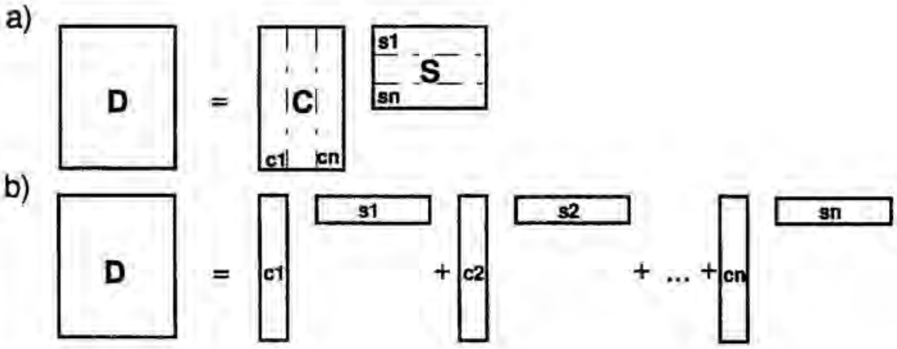


Figure 3.5.3.2.1. Two graphical views of the decomposition of a bilinear multicomponent data matrix. a) As the product of the matrices including the pure response profiles, b) as the sum of the products related to pure contributions of the mixture components.

In matrix notation, the general expression valid for all CR procedures is:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E}$$

where **D** *(r × c)* is the original data matrix, **C** *(r × n)* and **S** *(n × c)* are the matrices containing the pure response profiles related to the data variation in the row direction and in the column direction, respectively and **E** *(r × c)* is the error matrix, i.e., the residual variation of the data set that is not related to any chemical contribution. In a spectrometric titration, the **C** matrix contains the pure concentration profiles of all the absorbing species and the **S** matrix contains their related unit spectra (see Figure 3.5.3.2.2). For the sake of simplicity, **C** and **S** are referred to as concentration profile matrix and spectra matrix, though this does not mean that the applicability of CR methods is restricted to this kind of chemical data.

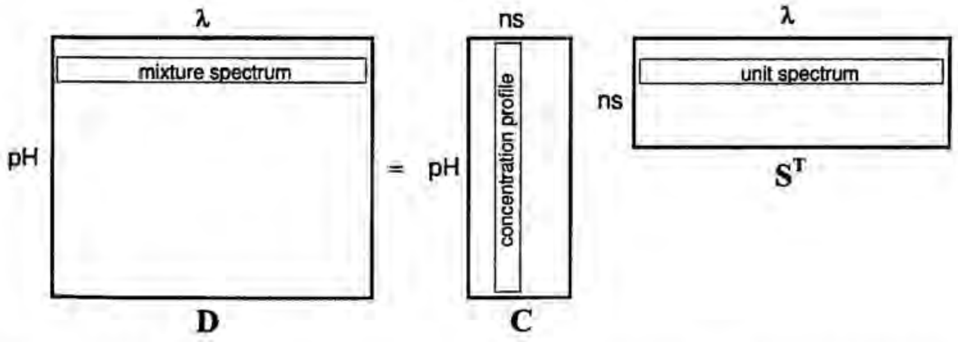


Figure 3.5.3.2.2. Output obtained from an acid-base spectrometric titration after ALS application. ns means number of absorbing species.

The decomposition of a single bilinear matrix is inherently affected by two sources of ambiguity: rotational ambiguity and intensity ambiguity (Lawton,1971), (Tauler,1995). Whereas the former accounts for the possibility of correctly reproducing the original data matrix by using **C** and **S** matrices containing linear combinations of the real profiles, the latter warns about the possibility of having profiles completely correlated with the real ones, though different in magnitude. In plain words, the correct reproduction of the original data matrix can be achieved by using response profiles differing in shape (rotational ambiguity) or in magnitude (intensity ambiguity) from the real ones.

The mathematical explanation of these two ambiguities is simple. The basic equation associated with CR methods, $\mathbf{D} = \mathbf{CS}^T$, can be easily transformed as follows:

$$\mathbf{D} = \mathbf{C}\,(\mathbf{T}\,\mathbf{T}^{-1})\,\mathbf{S}^T$$
$$\mathbf{D} = (\mathbf{CT})\,(\mathbf{T}^{-1}\,\mathbf{S}^T)$$
$$\mathbf{D} = \mathbf{C'}\,\mathbf{S'}^T$$

where $\mathbf{C'} = \mathbf{CT}$ and $\mathbf{S'}^T = (\mathbf{T}^{-1}\,\mathbf{S}^T)$ describe the **D** matrix as well as the true **C** and **S** matrices do, though **C'** and **S'** lack chemical sense. On the basis of the transformation shown in these equations, the mathematical formulation of the rotational ambiguity problem indicates that the possible solutions of a resolution method are as numerous as the **T** matrices can be, i.e., infinite. However, the inclusion of information related on the internal structure of the data (e.g. the presence of selective zones) and on their chemical properties in the resolution process often allows the suppression of this ambiguity or, at least, a large decrease in the number of feasible solutions.

When a system lacking rotational ambiguity is considered, the basic CR equation can still be rewritten as shown below:

$$\mathbf{D} = (1/k)\mathbf{C}\;k\mathbf{S}^T$$
$$\mathbf{D} = \mathbf{C'}\,\mathbf{S'}^T$$

where k is a scalar. The concentration profiles of the new $\mathbf{C'} = (1/k)\mathbf{C}$ matrix have the same shape as the real ones, but are k times smaller, whereas the spectra of the new $\mathbf{S'} = k\mathbf{S}$ matrix are shaped like the **S** spectra, though k times more intense. This ambiguity cannot be solved unless external information is introduced in the resolution process. Both rotational and intensity ambiguities are drastically diminished when several matrices are considered together.

The correct performance of the curve resolution methods depends strongly on the internal features of the data set being analyzed. Regardless of the quality of each method, the two following conditions must be fulfilled if the true concentration profile and spectrum of each compound in the data matrix is to be recovered (Manne,1995):

- The true concentration profile of an analyte can be recovered when all the compounds inside its concentration window are also present outside.
- The true spectrum of an analyte can also be recovered if its concentration window is not completely embedded inside the concentration window of a different compound.

The mathematical background of the CR methods is very diverse (Windig,1988), (Hamilton,1990), (Tauler,1995). Some of them use only abstract variables, such as window factor analysis, WFA (Malinowski,1992) or heuristic evolving latent projections, HELP (Kvalheim,1993), some others prefer the use of real variables, such as the simple-to-use interactive self-modelling analysis, SIMPLISMA (Windig,1991) or the orthogonal projection approach, OPA (Cuesta-Sánchez,1996), and some others transform iteratively abstract initial estimates into real solutions, such as iterative transformation target factor analysis, ITTFA (Gemperline,1984;1986), (Vandeginste,1985) or alternating least squares, ALS (Tauler,1995;1995b). The CR methods can also be classified according to the data sets to which they are applicable. Thus, the two-way resolution methods are applied to the resolution of single matrices, e.g. HELP, WFA, SIMPLISMA, ITTFA, ..., whereas the three-way resolution methods can deal with two (generalized rank annihilation method, GRAM (Sánchez,1986)) or more matrices together (trilinear decomposition, TLD (Sánchez,1990), parallel factor analysis, PARAFAC, (Bro,1996) and restricted Tucker models (Smilde,1994)). Methods like ALS are adapted to one or more matrices.

Among the different CR methods, the Alternating Least Squares method has been chosen to treat the data coming from the spectrometric titrations for the following reasons:

- **the iterative character.** Errors made in the initial estimates can be corrected in successive calculation steps.
- **the possible input of chemical information.** Some general features of the concentration profiles and/or of the spectra can be introduced in the resolution procedure by the application of suitable constraints.
- **the flexibility in the input of constraints.** Both the concentration profile matrix and the spectra matrix can be constrained. Each constraint is optionally applied to all, some or none of the species in the data set and departures more or less important from the complete fulfilment of a selected constraint are also admitted.
- **the adaptability to work with several matrices.** In contrast to most of the CR methods which are suitable for treating either one data matrix (two-way data) or a group of them (three-way data), ALS is equally applicable to both kinds of data set. When compared with other three-way resolution methods, ALS appears to be much less demanding since it does not limit the number of matrices to be treated together and only requires a common matrix direction (i.e., equal number of rows or equal number of columns).

3.5.3.2.1. Alternating Least Squares method (ALS).

The Alternating Least Squares method described in this text has been implemented by the research group of *Chemometrics and Solution Chemistry* at the Departament de Química Analítica de la Universitat de Barcelona and has been applied successfully to many diverse data sets (HPLC-DAD, spectrometric monitoring of kinetic, thermodynamic and structural transitions, voltammetric process monitoring, ...) with different degrees of complexity (multicomponent systems with variable overlap among the concentration profiles and instrumental responses) (Tauler,1993;1995;1995b), (Saurina,1995), (Lacorte,1995).

The resolution of a single data matrix by the ALS method includes the following steps:

1. Construction of an initial bilinear data matrix **D** with the experimental data.
2. Determination of the number of species.
3. Building a matrix of initial estimates.
4. Selection of the constraints to be applied in the iterative resolution process.
5. Application of the Alternating Least Squares optimization to obtain the definitive concentration profiles and spectra matrices.

The detailed explanation of these steps is given below. Some notes concerning the application of ALS to the particular case of polynucleotide acid-base spectrometric titrations (*NPT)* have also been included.

1. Construction of an initial bilinear data matrix D with the experimental data. The instrumental responses (e.g. spectra, voltammograms,...) must be structured to form a bilinear data matrix by sorting the data arrays one under the other as they are recorded to keep the information related to the evolution of the experiment.
NPT: the spectra must be set in the data matrix following a continuous increasing or decreasing pH sequence.

2. Determination of the number of species. Prior to the resolution of a mixture data matrix, the total number of independent and chemically meaningful contributions to this matrix must be known. As commented in previous sections, the assumption that the number of principal components equals the number of chemical contributions to the variation of the data set is usually invalid when dealing with evolutionary data sets; therefore, the graphical information provided by EFA and FSMW-EFA is generally more helpful than the application of more rigid statistical and/or empirical criteria to estimate the total number of species in a data set. However, the number decided in this step can be modified if the final results of the resolution procedure make no sense chemically.
NPT: in this context, the term species refers to absorbing species; therefore, all the compounds with different spectra, irrespective of whether they are chemical species or conformers, are considered equally.

3. Building a matrix of initial estimates. The iterative process to solve the equation $\mathbf{D} = \mathbf{CS}^T$ begins with two matrices: the initial data matrix, **D**, and a matrix of initial estimates (either concentration profiles, $\mathbf{C}_{in}$, or spectra, $\mathbf{S}_{in}$). The initial estimates are usually built by the EFA method, as shown in section 3.5.3.1.2, or by some other procedures that take into account the information in the original data set (e.g. FSMW-EFA, SIMPLISMA, needle algorithm (de Juan, 1996),...). Random estimates, much further from the true solutions, are not introduced since they slow down the iterative process and entail a higher risk of converging to local minima. Methods using these random estimates, such as the Alternating Regression, AR (Karjalainen,1989), run the iterative process several times with different initial estimates and provide a solution band for each response profile; the whole resolution process then becomes cumbersome and there is no gain of reliability in the final solutions.
NPT: the initial estimates are mostly matrices of concentration profiles because the overlap between profiles in this matrix direction is usually smaller than in the spectral direction.

4. Selection of the constraints to be applied in the iterative resolution process. In each iterative cycle, both **C** and **S** matrices can be modified according to certain constraints based on the internal structure of the data matrix **D** and on the chemical features of the concentration profiles and instrumental responses. A constraint updates in the response profiles the elements that do not fulfil a certain condition by some others that do. The use of all the constraints is optional and they can be applied either to all the profiles in the **C** and/or **S** matrices or only to some of them. Small departures from the conditions imposed by some constraints are also allowed.

The most essential constraint in the resolution process is **selectivity.** The selective zones in a data matrix are those regions related to only one species. Thus, a selective concentration region covers the range of the evolutionary variable (pH, time,...) where only one species is present, whereas a selective spectral region is the range of wavelengths where only one species absorbs. If a species has a selective concentration window, its spectrum can be perfectly recovered and if the selectivity is in the spectral direction, the concentration profile can be determined correctly. The presence of selective zones for all the species in a data matrix eliminates the rotational ambiguity and ensures the recovery of the real response profiles of the chemical system. EFA and FSMW-EFA are the most suitable techniques to detect and locate the selective zones of a system, as shown in previous sections. The selectivity for one species is respected by setting to zero the elements of all the other response profiles in the selective regions of this species.

In a wide sense, a chemical constraint is any systematic feature present in the concentration profiles or in the instrumental responses of a chemical system. The most frequently used are:

- **non-negativity.** This constraint forces the values in a profile to be equal to or greater than zero and is applied to all the concentration profiles and to some experimental responses, such as UV absorbances. The non-negativity constraint is applied by setting the negative values in a profile equal to zero or equal to an extremely small positive value.
- **unimodality.** This constraint allows the presence of only one maximum in each response profile. It is applied to chromatographic peaks, to the concentration profiles of some chemical reactions and to some peak-shaped instrumental responses, such as voltammograms. A possible implementation of the unimodal constraint consists of suppressing the non-unimodal part of a profile (i.e., the secondary maxima) and replacing the elements of this part by extremely small positive values.
- **closure.** This constraint is applied to closed reaction systems for which the sum of the concentrations of all the species involved in the reaction or some of them is known to be constant at each stage of the reaction. In each row of the concentration matrix, the elements related to the species involved in the closure are modified proportionally to make the sum of them equal to the closure constant.

The performance of the above constraints is graphically shown in the Figure 3.5.3.2.1.1.

NPT: selectivity is applied, when exists, in the concentration and the spectral direction. The concentration profiles are forced to fulfil the non-negativity, unimodality and closure constraints. The spectra coming from UV titrations are constrained to be positive, whereas the CD spectra do not observe this constraint.

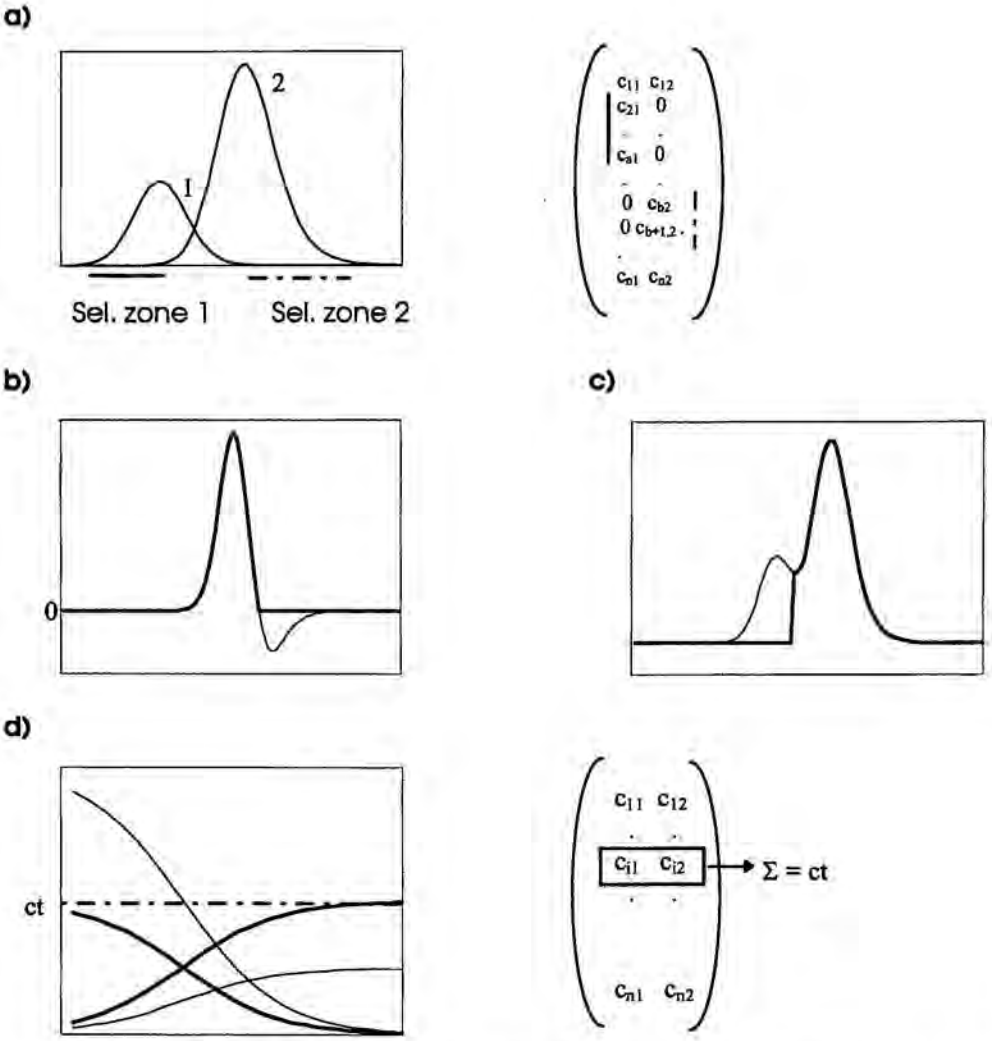


Figure 3.5.3.2.1.1. Performance of the constraints in the ALS method. a) Selectivity. The solid line and the dashed line indicate the selective zones for the compounds 1 and 2,respectively. On the right, the constrained matrix is shown. b) Non-negativity, c) unimodality and d) closure: the normal line and the thick line show a response profile before and after the application of the constraint, respectively. d) also presents the constrained concentration matrix.

5. Application of the Alternating Least Squares optimization to obtain the definitive concentration profiles and spectra matrices. Once the matrix of initial estimates is built and the constraints to be input in the resolution process have been decided, the iterative optimization starts. Each cycle of this optimization is as follows:

a) Calculation of the **S** matrix by using the original data matrix **D** and the matrix of initial estimates, $\mathbf{C_{in}}$.

$$\mathbf{D = C_{in}\,S^T}$$

$$\mathbf{S^T = (C_{in}{}^T C_{in})^{-1}\,C_{in}{}^T\,D}$$

or, in a shorter notation,

$$\mathbf{S^T = C_{in}{}^+\,D}$$

where $\mathbf{C_{in}{}^+ = (C_{in}{}^T C_{in})^{-1}\,C_{in}{}^T}$ is the pseudoinverse of the $\mathbf{C_{in}}$ matrix (Golub,).

b) Updating of the $\mathbf{S^T}$ matrix according to the selected constraints. $\mathbf{S^T \rightarrow S_r{}^T}$.

c) Calculation of a new concentration matrix **C** by using **D** and $\mathbf{S_r{}^T}$.

$$\mathbf{D = C\,S_r{}^T}$$

$$\mathbf{C = D\,S_r\,(S_r{}^T\,S_r)^{-1}}$$

$$\mathbf{C = D\,(S_r{}^T)^+}$$

where $\mathbf{(S_r{}^T)^+ = S_r\,(S_r{}^T\,S_r)^{-1}}$ is the pseudoinverse of the $\mathbf{S_r{}^T}$ matrix.

d) Updating of the **C** matrix according to the suitable constraints. $\mathbf{C \rightarrow C_r}$.

e) Reproduction of the original data matrix by using the constrained matrices $\mathbf{C_r}$ and $\mathbf{S_r{}^T}$.

$$\mathbf{D^* = C_r\,S_r{}^T}$$

Analogous steps involving the calculation of the constrained $\mathbf{C_r}$ and $\mathbf{S_r{}^T}$ matrices in this order are followed when the matrix of initial estimates contains spectra. When the reproduced matrix **D*** and the original data matrix **D** are similar enough, the iterative optimization is concluded and $\mathbf{C_r}$ and $\mathbf{S_r{}^T}$ are taken as final solutions. If this does not happen, a new iterative cycle is started with the matrix $\mathbf{C_r}$ playing the same role as the initial estimate $\mathbf{C_{in}}$ in the first iteration. The iterative process is finished when convergence is achieved or when a preselected number of iterative cycles is exceeded. The convergence criterion is here defined in terms of the change in relative standard deviation of the residuals between two consecutive cycles. When this change is smaller than a selected value, the optimization process is finished.

NPT: the concentration profiles and spectra obtained with ALS provide complete information about the evolution of all the absorbing species present in the acid-base equilibria and about their spatial structure. Nevertheless, the task of identifying which transitions take place and which conformations are involved in them is beyond the scope of the method and depends only on the previous knowledge about the process being analyzed.

The ALS method is easily adapted to handle three-way data sets, i.e., formed by more than one data matrix. The working procedure of ALS only requires one order in common between the matrices to be treated together, as shown in Figure 3.5.3.2.1.2. This means that the number of rows and/or the number of columns of the different matrices must be the same and the variables in the common order must have the same chemical meaning. The three-way data set formed by single matrices sharing the same number of rows can be written as an augmented row-wise matrix, where the single matrices are placed one beside each other. Analogously, single matrices sharing the same number of columns form an augmented column-wise data matrix where the single matrices are placed one on top of each other.

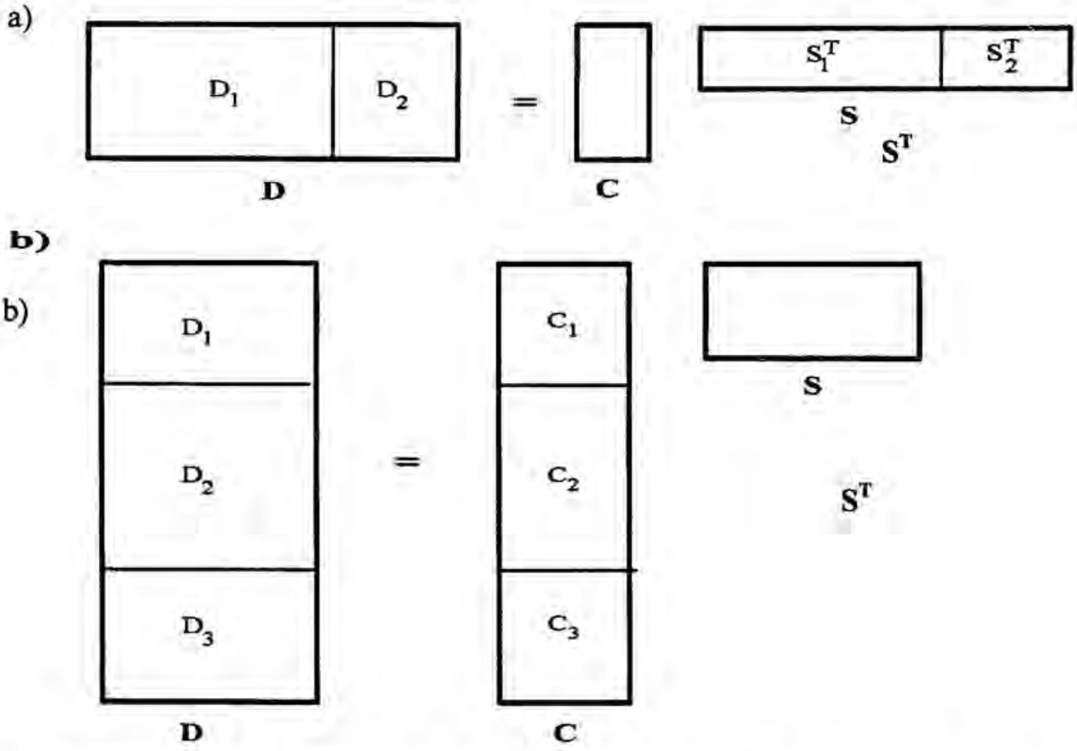


Figure 3.5.3.2.1.2. ALS application to three-way data sets. a) row-wise augmented data matrix and b) column-wise augmented data matrix.

The resolution results obtained for a data matrix included in a three-way data set will always be better or, at least, equal to those obtained when this matrix is handled alone. The ambiguities related to the decomposition of a single bilinear data matrix are usually suppressed or, at least, greatly diminished by the inclusion of further information and additional constraints when several matrices are treated together. The ALS method follows basically the same operating procedure presented for a single data matrix when dealing with three-way data sets. However, slight differences related to the different steps described on previous pages deserve comment. The explanations given hereafter are related to a column-wise data set because this is the kind of three-way data set obtained from the simultaneous treatment of several polynucleotide spectrometric titrations. Analogous reasoning can be applied to the treatment of row-wise data matrices. *NPT* are also included when necessary.

1. Construction of the initial three-way data set D from the experimental data. The initial column-wise augmented data matrix is built by placing the single matrices related to each experiment one on top of each other. The order in which these matrices are appended does not affect the final resolution results.

NPT: the spectrometric titrations only share the spectral order in common (i.e., the wavelength range at which the UV or CD spectra have been recorded). The titrations performed with the same spectrometric technique are treated together.

2. Determination of the number of species. The number of species is independently determined for each single matrix in the three-way data set.

3. Building a matrix of initial estimates. In the resolution of a column-wise augmented data matrix, the initial estimates can be either a single spectra matrix or a column-wise concentration matrix. The column-wise concentration matrix is built by placing the initial concentration estimates obtained for each single data matrix in the three-way data set one on top of each other. The appended initial estimates must be sorted as the initial data matrices are in **D** and must keep a correct correspondence of species, i.e., each column in the augmented concentration matrix must be formed by appended concentration profiles related to the same species. When no prior information about the identity of the species in the different data matrices is available, the correct correspondence of species can be achieved by looking at the resolution results of each single matrix. The matrices of initial estimates must have the same number of columns to be appended. Since their related original data matrices can present different number of species, the columns related to absent species in the initial estimate of a single data matrix are filled with zeros, as shown in Figure 3.5.3.2.1.3.

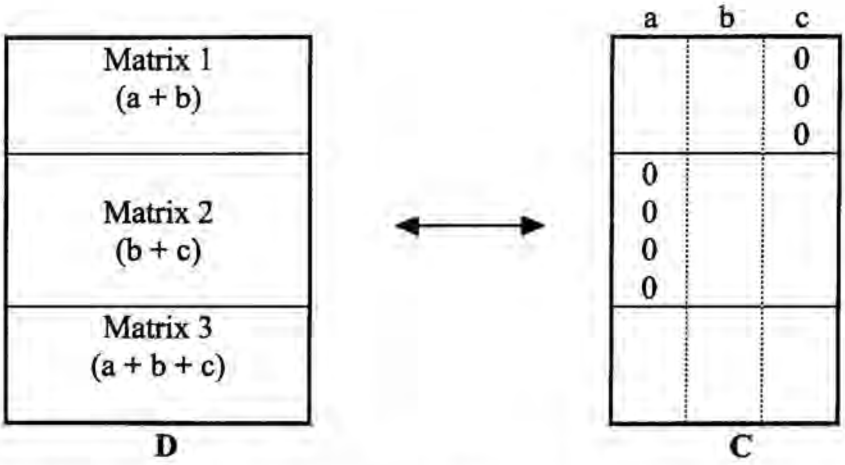


Figure 3.5.3.2.1.3. Column-wise initial estimate (C) of a three-way data set (D). a, b and c are the species present in the data set.

4. Selection of the constraints to be applied in the iterative resolution process. The same constraints applied in the resolution of a single data matrix can be applied to three-way data sets. Selectivity and non-negativity affect the spectrum and the augmented concentration profile of each species, whereas unimodality is applied separately to each of the single profiles appended to form the augmented concentration profile. The closure constraint operates applying the corresponding closure constant to each of the single matrices in the column-wise concentration matrix.

Apart from these constraints, some others are exclusively applied to three-way data sets, namely:

- **common species have the same spectrum in all the appended data matrices.** This is a constraint inherently present due to the decomposition process of column-wise data matrices (see Figure 3.5.3.2.1.2.). The possibility to have only one spectrum for each species in all the appended data matrices makes the rotational ambiguity less severe than in the case of single matrices.

NPT: this constraint can be perfectly assumed since the titrations treated together have been performed in the same experimental conditions (T, ionic strength,...) and hence, no variations in the spectrum of a species can be expected among experiments.

- **common species have concentration profiles with equal shape in all data matrices.** In contrast to the previous constraint, this one can be optionally applied. The ALS decomposition of a three-way data set does not involve any assumption as to how the concentration profiles of the species change in the matrices analyzed together. This constraint can only be applied to data sets formed by matrices with two orders in common that are intrinsically trilinear (i.e., owing to the features of the chemical process, the shape of the concentration profiles of common species is known to be equal among experiments). When this condition is fulfilled, the resolution process gives unique solutions (i.e., not affected by rotational ambiguity) even in the absence of selectivity. Some methods based on the comparison of the rank between matrices augmented in different directions have been proposed to detect the presence of trilinear structure in a three-way data set (Tauler,1995).

  The application of this constraint (also called trilinearity) in the ALS method is performed separately on the concentration profile of each species. Thus, the single profiles forming the augmented concentration profile of a certain species are placed one beside each other to form a new data matrix. PCA is performed on this matrix. If the system is trilinear, the score vector related to the first PC must show the real shape of the concentration profile and the rest of PCs must be related to noise contributions. The loadings related to the first PC are scaling factors accounting for the species concentration level in the different appended matrices. Therefore, the new single profiles will be calculated as the product of the score vector by their corresponding scaling factor. The constrained single profiles are finally appended to form the new augmented concentration profile. All this process is graphically shown in Figure 3.5.3.2.1.4.

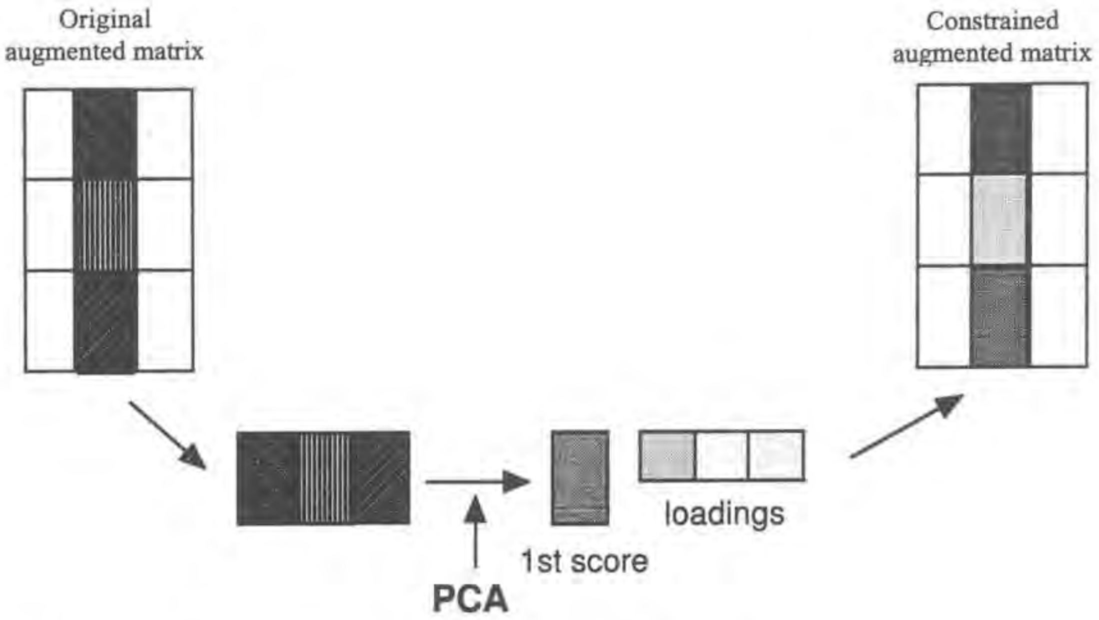


Figure 3.5.3.2.1.4. Application of the trilinear constraint in the ALS method.

- **correspondence between species.** In each single matrix of a three-way data set, the elements in the concentration profiles of absent species are set equal to zero after each iterative cycle. This constraint is not present in other three-way resolution methods, such as GRAM or TLD, where all the species are modelled in each single data matrix, even if some of them are known to be absent.

NPT: the same constraints used for the treatment of a single titration are imposed in the simultaneous treatment of several experiments. No trilinearity can be applied since the number of rows (spectra) in each experiment is different. Anyway, even if the experiments were performed in equal conditions, the inclusion of trilinearity would not be very helpful. Two data matrices coming from acid-base titrations of the same system performed in the same conditions only differ in a scaling factor related to the different concentration levels used in both experiments. This means that the input of a new matrix does not furnish any new information to be used in the resolution of the data set.

5. Application of the Alternating Least Squares optimization to obtain the definitive concentration profiles and spectra matrices. The optimization process follows the same steps explained before. For a three-way data set, the whole augmented concentration matrix is treated in the same way as a single concentration matrix in the resolution of a single data matrix.

The use of three-way resolution methods can also provide quantitative information about the data set. Since the spectrum of one species is common to all the appended data matrices, the area of the single concentration profiles of this species is scaled according to the concentration level of the species in each single data matrix. Thus, the relative concentration of a particular species in one of the appended matrices can be obtained from the quotient between the area of its concentration profile and the area related to the concentration profile of the same species in a matrix taken as reference.

Either dealing with single data matrices or with three-way data sets, the quality of the ALS results is evaluated by calculating the lack of fit, expressed as follows:

$$\text{lack of fit} = \sqrt{\frac{\sum_{ij}(d_{ij} - d_{ij}^{*})^2}{\sum_{ij} d_{ij}^2}}$$

where d_{ij} are the data in the original data set and d_{ij}^{*} are the same data reproduced by using the matrices of concentration profiles and spectra obtained with the ALS method.

II. Results

CHAPTER 4.

ON THE MICROSCOPIC DESCRIPTION OF SOLVENT SPACE

Solvent classification based on solvatochromic parameters: a comparison with the Snyder approach

Anna de Juan *, Gemma Fonrodona, Enric Casassas
Barcelona, Spain

Solvents play a major role in many chemical processes. Their effect is closely related to the nature and extent of solute/solvent interactions, developed locally in the immediate vicinity of the solutes. A new generation of empirical parameters helps to describe the solvent features in these microscopic environments. Their adequacy for the establishment of linear free energy relationships (LFER) has been widely proved. A classification scheme based on the microscopic properties of solvents is proposed, to provide a clearer insight into the solvent-space structure in terms of the similarity of solute/solvent interactions and to help chemists in the choice of a suitable solvent for each purpose. The Kamlet and Taft solvatochromic parameters α, β and π^* have been selected as microscopic solvent descriptors, and solvent groups have been designed according to the results obtained from the application of several multivariate clustering techniques. The present scheme is compared with Snyder's triangle, an earlier solvent representation used extensively in chromatography. Analogies and differences which arise from the field of application and from the solvent descriptors used in each classification scheme are discussed.

1. Introduction

For centuries chemists have been challenged to explain the structure of solvents and their interaction with solutes. From early times, changes in the working media were seen to cause alterations in the extent and direction of chemical equilibria, in the order and rate of reaction kinetics or in spectroscopic behaviour, to mention only a few examples, thus confirming the active role of solvents in the evolution of chemical processes [1].

The physical constants of solvents (i.e. their melting and boiling points, refraction index, dielectric constant, etc.) were for many years the only available descriptors for these substances. However, inconsistencies in the behaviour of solutes were often found in regard to the properties of the

*Corresponding author.

0165-9936/97/$17.00
PII S0165-9936(96)00079-9

Table 1
Solvatochromic parameters for the selected solvent set

Solvent		α	β	π^*
1	Diisopropyl ether	0.00	0.49	0.27
2	Di-*n*-butyl ether	0.00	0.46	0.24
3	Diethyl ether	0.00	0.47	0.27
4	Dioxane	0.00	0.37	0.55
5	Tetrahydrofuran	0.00	0.55	0.58
6	Anisole	0.00	0.22	0.73
7	Dibenzyl ether	0.00	0.41	0.80
8	Diphenyl ether	0.00	0.13	0.66
9	Fenetole	0.00	0.20	0.69
10	2-Butanone	0.06	0.48	0.67
11	Acetone	0.08	0.48	0.71
12	Ethyl acetate	0.00	0.45	0.55
13	Ethyl benzoate	0.00	0.41	0.74
14	Propylene carbonate	0.00	0.40	0.83
15	Dimethylacetamide	0.00	0.76	0.88
16	Dimethylformamide	0.00	0.69	0.88
17	N-Methylpyrrolidone	0.00	0.77	0.92
18	Tetramethylurea	0.00	0.80	0.83
19	Triethylamine	0.00	0.71	0.14
20	Dimethyl sulfoxide	0.00	0.76	1.00
21	Hexamethylphosphorotriamide	0.00	1.05	0.87
22	Nitrobenzene	0.00	0.39	1.01
23	Benzonitrile	0.00	0.41	0.90
24	Acetonitrile	0.19	0.31	0.75
25	Pyridine	0.00	0.64	0.87
26	2,6-Dimethylpyridine	0.00	0.76	0.80
27	Quinoline	0.00	0.64	0.92
28	Toluene	0.00	0.11	0.54
29	Benzene	0.00	0.10	0.59
30	Chlorobenzene	0.00	0.07	0.71
31	Bromobenzene	0.00	0.06	0.79
32	Carbon tetrachloride	0.00	0.00	0.28
33	1,2-Dichloroethane	0.00	0.00	0.81
34	Methylene chloride	0.30	0.00	0.82
35	Chloroform	0.44	0.00	0.58
36	*tert*-Butanol	0.68	1.01	0.41
37	Isopropanol	0.76	0.95	0.48
38	*n*-Butanol	0.79	0.88	0.47
39	Ethanol	0.83	0.77	0.54
40	Methanol	0.93	0.62	0.60
41	Ethylene glycol	0.90	0.52	0.92
42	Water	1.17	0.18	1.09

solvents used. These apparent contradictions revealed that solvents should not be considered as macroscopic continua, but as dynamic structures having molecules that interact differently with each other and with solutes. The macroscopic parameters can thus be used as descriptors of the bulk solvent but are unable to provide a reliable picture of the solvent structure around the solutes, in their solvation sphere. Therefore, a new generation of parameters was needed to define the features of the cybotactic zone of solutes. The monitoring of a solvent-sensitive reference process is the common basis of all the methods devoted to establishing empirical scales of microscopic parameters focused on the quantitative description of solute–solvent interactions. In all these scales, the reference solute behaves as a probe within the solvation shell, reflecting changes in the surrounding solvent through variations in its absorption spectra [2–6] or in some well described thermodynamic [7,8] or kinetic processes [9,10].

The uniparametric scales [2,3,9] try to incorporate all solute–solvent interactions in a single parameter, which is often said to be a polarity parameter (the term polarity being interpreted as a rough concept involving all specific and non-specific interactions between solvent and solute [1]). In contrast, the multiparametric approaches [4–6,10,11] associate each kind of solute–solvent interaction (e.g., hydrogen-bonding or polarizability) with a separate parameter, and all are necessary to give a global picture of the solvent. The uniparametric and multiparametric approaches provide complementary information, focusing on the extent and the nature of the solute–solvent interactions, respectively.

The values of $E_T(30)$ [3] and the group formed by α, β and π^* [4–6] are the parameters most used in the uniparametric and multiparametric approaches. All of them can be measured from spectral shifts in the absorption bands of some reference solutes – hence the name, solvatochromic parameters. The quantity $E_T(30)$ has been defined as a polarity parameter and α, β and π^* are assumed to represent the hydrogen-bond acidity, hydrogen-bond basicity and polarity-polarizability, respectively. Several studies have shown a relationship between the two groups of parameters in pure [12] and mixed [13,14] solvents, $E_T(30)$ being a linear combination of π^* and α.

The present work is devoted to developing a solvent-classification scheme based on the similarity of the microscopic properties of the solvents, represented here by their α, β and π^* values. In order to avoid the biased criteria of the chemist in the solvent-grouping procedure, several clustering methods (hierarchical and non-hierarchical) have been employed. The diversity of their mathematical backgrounds allows more solid conclusions to be drawn, unaffected by specific trends which could

	$P'x_e$	$P'x_d$	$P'x_n$	ß	α	π*
$P'x_e$	*1.0000*	*0.6087*	*0.6117*	*0.7169*	*0.5607*	*0.4032*
$P'x_d$	*0.6087*	*1.0000*	*0.7151*	0.1528	*0.3476*	*0.6889*
$P'x_n$	*0.6117*	*0.7151*	*1.0000*	0.2108	-0.0086	*0.7653*
ß	*0.7169*	0.1528	0.2108	*1.0000*	0.2919	0.0338
α	*0.5607*	*0.3476*	-0.0086	0.2919	*1.0000*	0.0058
π*	*0.4032*	*0.6889*	*0.7653*	0.0338	0.0058	*1.0000*

Fig. 1. Correlation matrix for the selected microscopic parameters, evaluated from the selected solvent set. Double line: correlation coefficients among the Snyder's parameters. Thick line: correlation coefficients among the solvatochromic parameters. Single line: correlation coefficients among the Snyder's parameters and the solvatochromic parameters. The figures in italics are the significant correlation coefficients, according to the results obtained through the application of a suitable statistical test [33].

arise from the application of a concrete clustering technique [15].

2. Data set and clustering procedures

The original data matrix to be clustered is made up of 42 pure solvents (objects) characterized by their α, β and π* values (variables), as shown in Table 1. The object selection has been performed taking into account the size and representativeness of the pool of solvents.

The selection of the variables has been carried out bearing in mind the microscopic character of the solvent scheme; therefore, in contrast with other approaches [16,17], no bulk solvent properties have been included. Empirical polarity parameters such as $E_T(30)$, which involve many solute–solvent interactions, have also been discarded since most of them can be described by linear combinations of more specific microscopic parameters. Within this latter group, the selection of the solvatochromic parameters α, β and π* proposed by Kamlet and Taft is justified for the following reasons.

- They are easily determined experimentally for pure and mixed solvents, using spectroscopic measurements.
- They have widespread use and proven ability to accurately describe variations in a variety of solvent-dependent processes through LFER [18–22].
- Their optimal statistical and chemical features are reflected in the total decorrelation among them (as shown in Fig. 1), and in the complete independence of their chemical meaning.
- They are on similar scales, which permits cluster analysis to be performed on the raw data and, as a consequence, allows straightforward interpretation of the final results.

The present three-dimensional solvent space permits display of the data set in a plot, the *X*, *Y* and *Z* axes being the α, β and π* solvatochromic parameters (see Fig. 2). Several groups of objects arise from the space structure, but their degree of tightness is quite variable and the assignment of some objects to a concrete group is quite ambiguous. Although the human eye is one of the best pattern recognizers, a simple visual inspection – even assisted by a powerful 3D graphical spinning device – would not be rigorous enough to establish consistent clusters.

The necessary clustering techniques have been selected according to the data set structure and to the information requested. The combination of a display method with one or several clustering techniques is a suitable strategy for any grouping procedure [15]. Principal component analysis (PCA) is the most recommended visualization technique. This reduces the redundant original variable space to an optimal lower dimensional abstract space, whose variables contain uncorrelated information. This dimension decrease allows the objects to be mapped in a two- or three-dimensional space, and the coordinates of the objects in this new abstract space, the so-called scores, are often used as input data in many multivariate clustering techniques. However, the two main benefits provided by PCA (i.e. a reduction in the space dimension, and decorrelation of the variables) are not necessary in this example of solvents since the raw data are three-dimensional and the original variables have proven

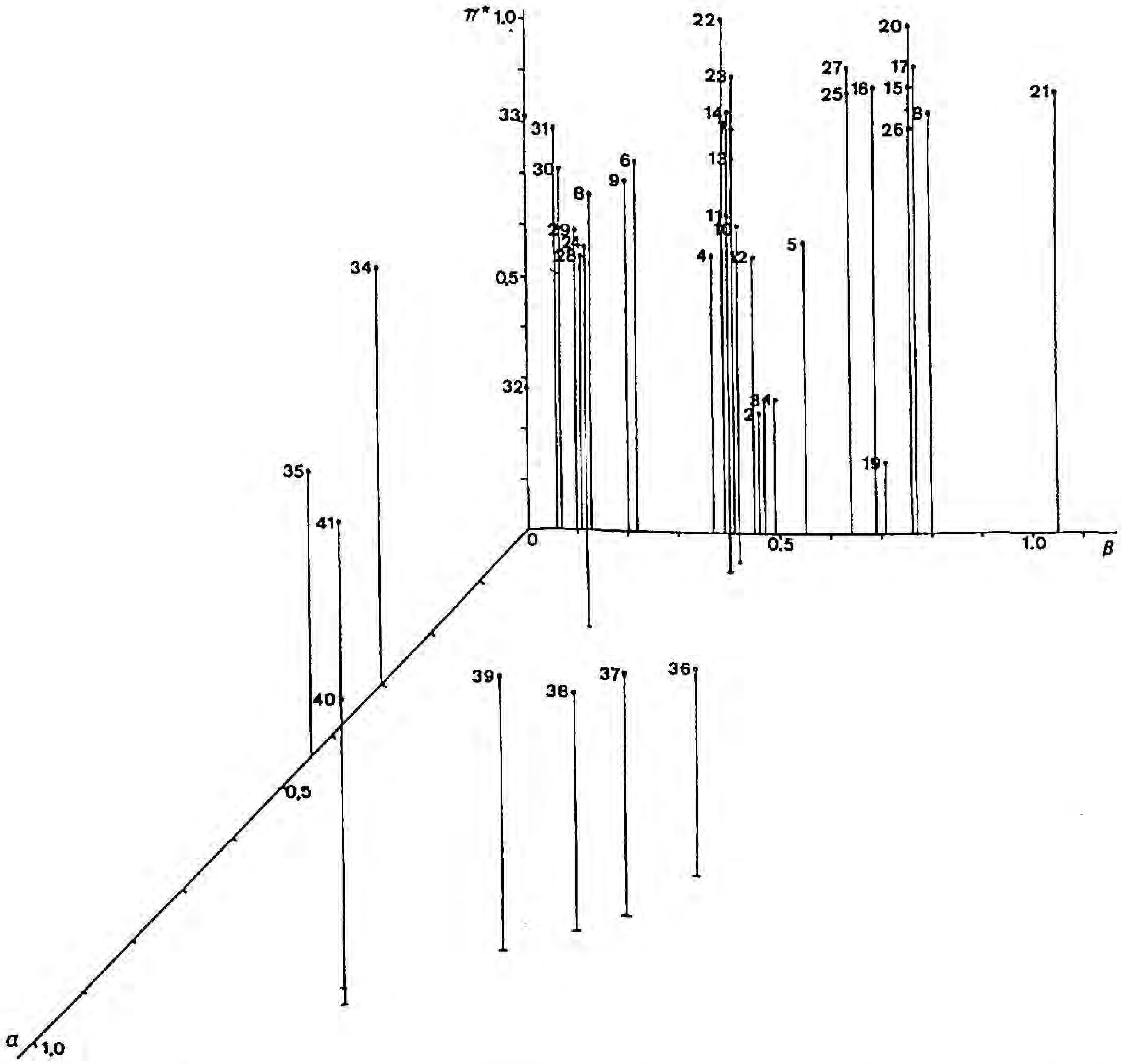


Fig. 2. Plot of the selected solvent set in the three-dimensional solvatochromic space. The data labels indicate the solvent numbers, according to Table 1.

not to be significantly correlated. Despite this reasoning, a PCA has been carried out on the autoscaled data and the two first PCs have been found to account for only 76% of the total variation in the data. This confirms the inability of PCA to reduce the dimensions of the solvent space, owing to the non-redundancy of the original variables.

The character of the above classification scheme, which is devoted more to finding out the general features within the different solvent groups than to obtaining information about individuals, has excluded the use of clustering methods based on the similarity among pairs of objects (i.e. single [24] and complete linkage [25]). Group average (GA) [26], Ward's method (WM) [27] and mode analysis (MA) [28] have been chosen as our hierarchical methods. All work with grouping criteria is based on properties either of the whole clusters (GA and WM) or on characteristics of the space regions (MA). The less common MA method works by mapping the density of the original data space. All the objects are first sorted in decreasing order of density and then afterwards introduced, one by one, in the clustering procedure. A new object can then become either the nucleus of a new cluster or a member of an existing group; owing to its spatial location a new object can also cause the fusion of two previously separated clusters. In the whole process of object introduction a maximum number of clusters is reached. This maximum is meant as being the lowest 'natural' classification level of the data set.

Complementary information is also supplied by using non-hierarchical procedures, such as RELOCATE [29], based on the successive steps of optimization of an initial K-clustering, linkage of the two nearest clusters, further optimization of the $K-1$ clustering, and so forth. The selection of the initial clustering can affect the final results, leading the final solutions to a local optimum. Therefore, the results presented in the present work come from two different initial configurations: the first, completely random, and the second, whose first K clusters were obtained through the application of Ward's method.

The use of fuzzy clustering procedures, although recommended for sparse distributions of objects, has been avoided. These methods assign each

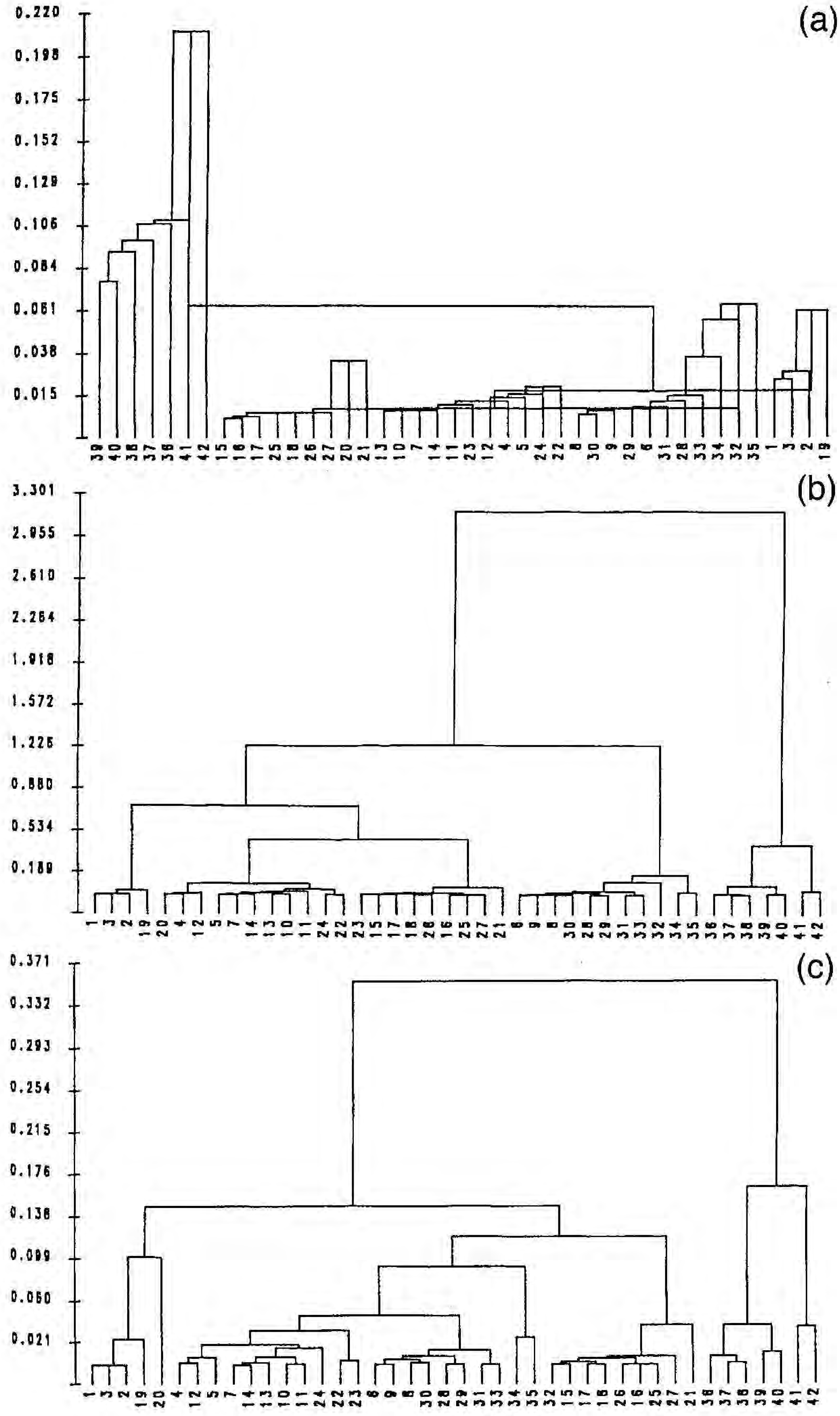


Fig. 3. Dendrograms obtained from the data set included in Fig. 1, using the solvatochromic parameters as clustering variables according to hierarchical mode analysis (a), Ward's method (b) and group average method (c).

object partially to several groups and consequently organise the original population in overlapped clusters. Since the present work is focused on clarifying the structure of the microscopic solvent space, methods providing separate groups of solvents have been considered more appropriate for this purpose.

All the selected clustering techniques have been applied as implemented in the CLUSTAN package [30]. The Euclidean distance has been adopted as the similarity measurement.

3. Results and discussion

The criterion implemented in the hierarchical mode analysis (MA) [28] has been adopted to decide the number of significant solvent classes.

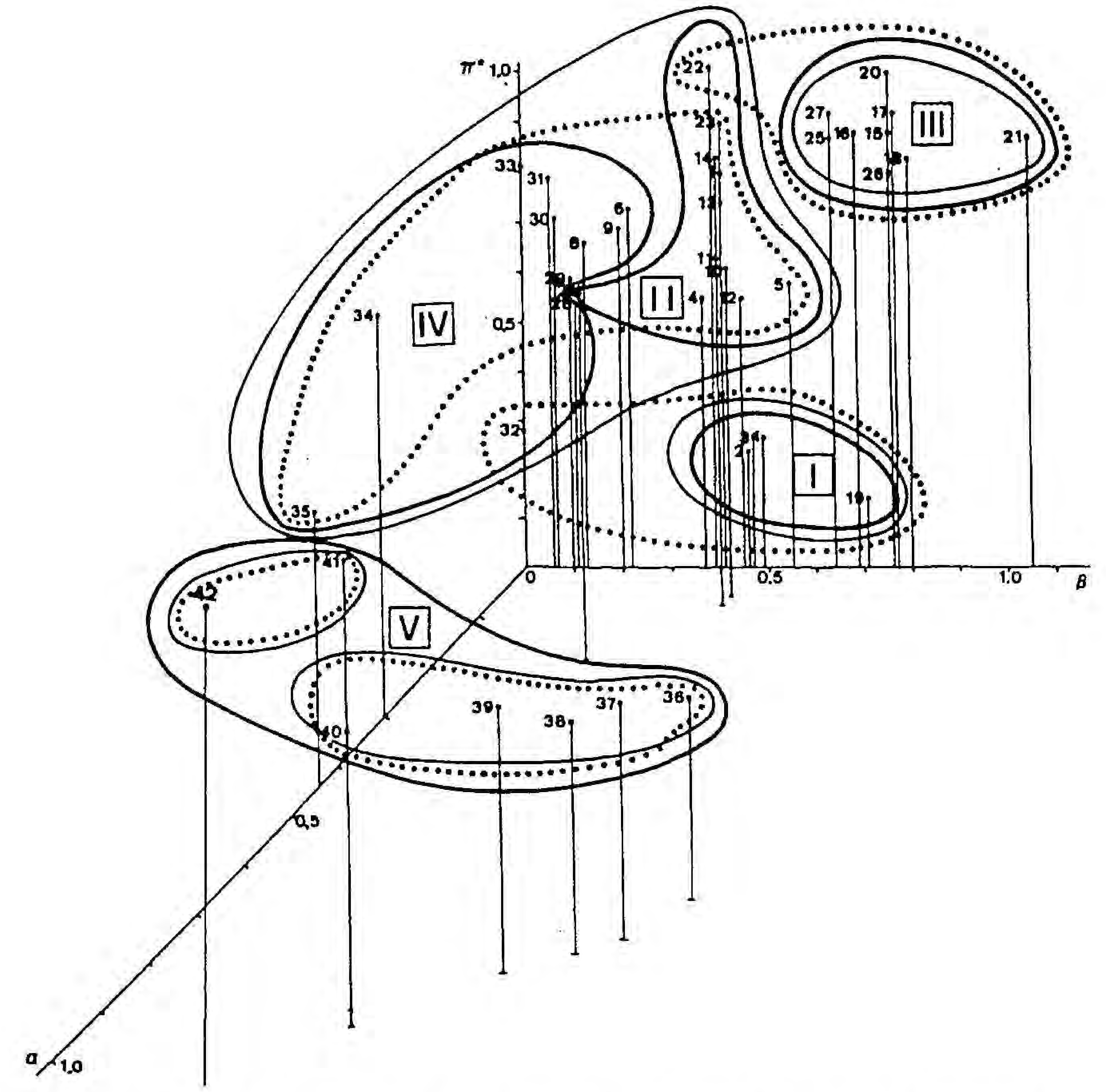


Fig. 4. Clusters plot obtained through the application of several clustering techniques. Thick lines: clusters obtained using hierarchical mode analysis and Ward's method. Single lines: clusters obtained with the RELOCATE method, for all the initial clustering distributions used. Dotted lines: clusters obtained from the group-average method. The data labels indicate the solvent number, according to Fig. 1. The clusters adopted in the solvent scheme, those circled with thick lines, are labelled with roman numerals as described in the text.

Thus, the solvent data set has been found to enclose five 'natural' clusters.

The dendrograms in Fig. 3a–c show graphically the clustering results related to the hierarchical MA, WM and GA methods, respectively. When the number of clusters is equal to five, the correspondence among the results coming from all the applied methods can clearly be examined through observation of Fig. 4, where the solvent classes obtained according to the different procedures have been drawn on the 3D solvents space. Note that the shape of the clusters has no mathematical meaning: the divisional lines are only useful for distinguishing which objects belong to each cluster. Fig. 4 shows two main clustering trends: the first, arising from the WM and the MA methods, and the second coming from GA and RELOCATE procedures.

Within the former trend described above, two clustering techniques with different mathematical backgrounds give rise to the same clusters. When the latter of trends is analyzed, the total agreement of the results coming from the RELOCATE procedure applied to different initial distributions once again proves the robustness of the data structure. Furthermore, the great similarity between the groups obtained through the use of the GA and RELOCATE procedures shows that the application of hierarchical or non-hierarchical methods is not responsible for the differences in the two detected clustering tendencies.

The results arising from Ward's method and hierarchical mode analyses have been selected finally to propose the definitive solvent classification scheme because both methods give rise to 'natural' clusters, without any constraint in shape, matching the usual irregular form of real data clusters. Moreover, the degree of discrimination of GA and MA is greater, and both tend to give similarly sized clusters. Indeed, the effect of very isolated objects, such as water, in the clustering procedures is much greater for the group average and RELOCATE methods. For these, the presence of such objects leads to the emergence of clusters which are too small and, as a consequence, to poorer differentiation of some other object groups which are closer among them but tighter and clearly separated.

Nevertheless, clusters equal or quite similar to those obtained through the application of Ward's method can also be found by using group average and RELOCATE methods in some other splitting levels, i.e. when the total number of clusters (K) is other than five.

Thus, the definitive solvent classification scheme includes the following clusters.

Cluster I. Slightly basic solvents (electron pair donors) with low polarity resulting from their aliphatic chains. This group contains aliphatic ethers and substituted aliphatic amines.

Cluster II. Aprotic polar solvents, all are relatively basic and moderately or highly polar. It includes aliphatic cyclic ethers, solvents containing the carbonyl functional group (esters and ketones) and nitriles.

Cluster III. Strongly basic and strongly polar solvents. Pyridines, small amides, sulfoxides, ureas and phosphoramides belong to this group.

Cluster IV. More heterogeneous than previous clusters, it includes relatively polar solvents, with a generally low tendency to form hydrogen bonds. Within this cluster, the next separation step splits the solvents 34 and 35 (the only moderate hydrogen-bond donors) from the rest, yielding two much tighter subclusters (A and B) with solvents having more similar features. Subcluster A contains aromatic compounds (ethers, hydrocarbons and halogenated species) and apolar aliphatic halogenated hydrocarbons. Subcluster B has polyhalogenated polar aliphatic hydrocarbons, whose heteroatoms cause the inductive effect responsible for the hydrogen-bond donor interaction.

Cluster V. This solvent class, formed by amphiprotic solvents (alcohols and water) with marked hydrogen-bond properties (donor and acceptor), is the most disperse and differentiated group. In successive splittings, solvents 41 and 42 (glycol and water, respectively) are separated from the other alcohols owing to their greater association ability.

Cluster VI. Although their elements are not represented in the solvent space, this group would be formed by the aliphatic hydrocarbons with null solute–solvent interactions, that is to say, with α, β and π* equal to zero.

All the relevant information concerning this solvent scheme (i.e. members of each solvent class, central coordinates of each cluster) is collected in Table 2.

The hydrogen-bond basicity appears to be the most discriminating parameter in the organization of the solvent scheme. Thus, more significant differences can be detected among the β-central coordinates of the groups than among the values of the α and π* parameters. This is specially noticeable in clusters II and IV where the different basicity of their members is the only element of variation between the two solvent classes. The effect of some other solvent features, such as the overall ability of interaction, is also identified as a cause of splitting among groups. In this sense, both clusters I and III include solvents which can develop polar and basic interactions with solutes; the only difference in this case is that the intensity of solute–solvent interactions in grcup III is larger than in group I. Cluster V confirms that the hydrogen-bond-donor ability, although less commonly present, is also a strong differentiating factor among solvents. A feature common to all the groups established in this solvent scheme is the variety in size, functional groups and spatial structure of the members included in each cluster. This internal diversity within groups could be considered as one of the most remarkable characteristics of this solvent classification since it reveals the hidden similar behav-

Table 2
Cluster information according to the results obtained using α, β and π* as clustering variables

Cluster number		Solvent numbers	Cluster central coordinates		
			α	β	π*
1		1, 2, 3, 19	0.000	0.355	0.265
2		4, 5, 7, 10, 11, 12, 13, 14, 22, 23, 24	0.030	0.424	0.735
3		15, 16, 17, 18, 20, 21, 25, 26, 27	0.000	0.763	0.885
4	A	6, 8, 9, 28, 29, 30, 31, 32, 33	0.067	0.081	0.654
	B	34, 35			
5		36, 37, 38, 39, 40, 41, 42	0.866	0.745	0.644

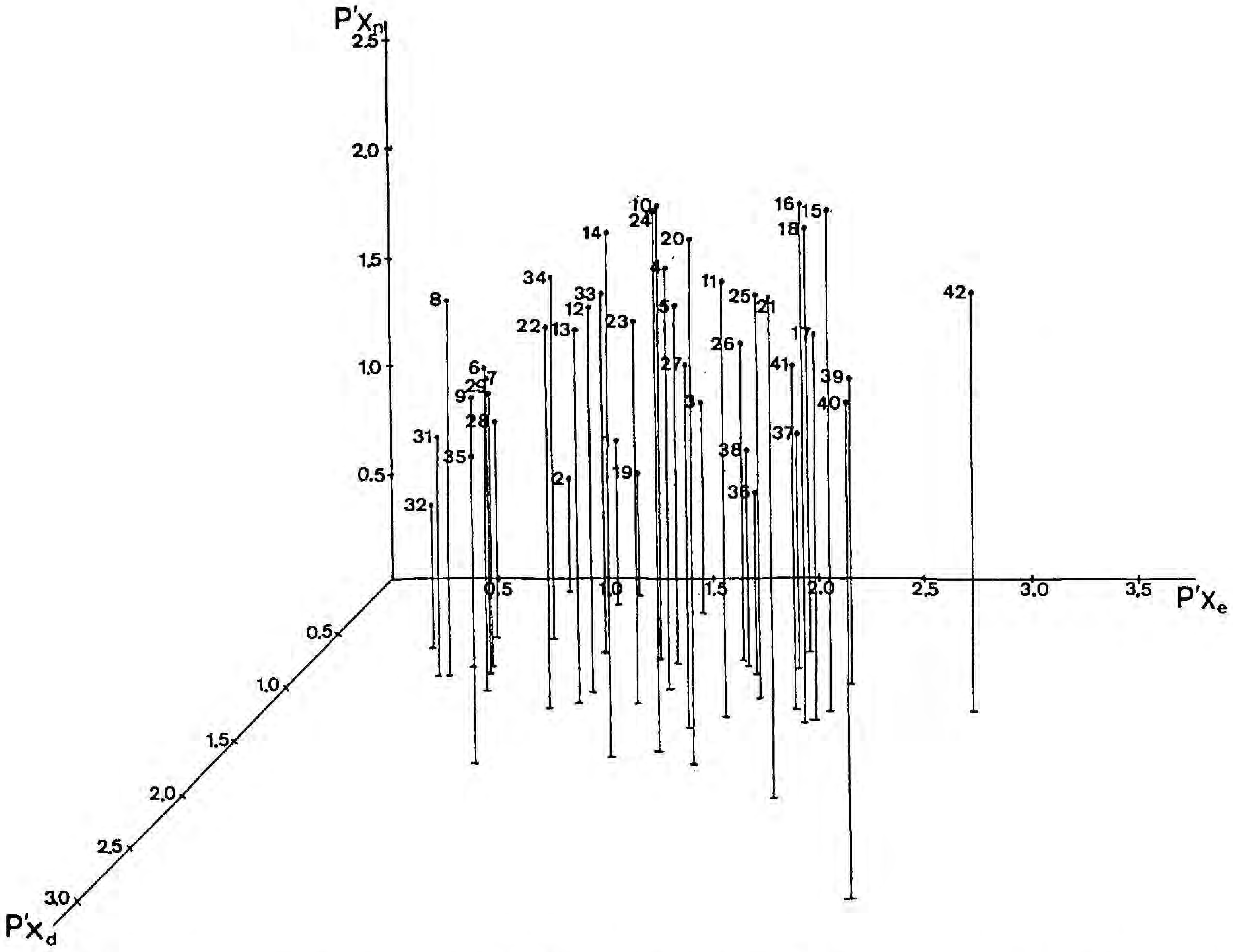


Fig. 5. Plot of the selected solvent set in the three-dimensional Snyder's space. The data labels indicate the solvent number, according to Table 3.

iour in terms of solute–solvent interactions among solvents traditionally classified as being further apart. This fact offers a wide range of possibilities when a solvent must be selected for a certain chemical process, since limitations produced by steric hindrance or by specific interactions with a certain functional group can be overcome. Additional problems related to toxicity, flammability or some other undesirable properties of certain solvents can also be avoided by the selection of suitable alternative compounds.

4. Comparison with Snyder's approach

Before the extended use of most of the microscopic solvent parameters, Snyder proposed one of the most popular and widely used solvent classifications: the solvent-selectivity triangle, which is known specially in the chromatographic field [31]. Snyder chose the appropriately corrected logarithms of certain gas–liquid distribution constants of some reference solutes, namely dioxane $[\log(K_g'')_d]$, ethanol $[\log(K_g'')_e]$ and nitromethane $[\log(K_g'')_n]$ as solvent descriptors related to the hydrogen-bond acidity, hydrogen-bond basicity and polarity, respectively. For each solvent, the polarity parameter P' is defined as the sum of these three logarithms. The terms $x_i = \log(K_g'')_i / P'$ (the subscript i referring to the reference solutes above) represent the weight of each of the solute–solvent interactions associated with dioxane, ethanol or nitromethane with respect to the global interaction ability of the solvent. As x_i are relative contributions, $x_e + x_d + x_n = 1$, and this is the base of the solvent-selectivity triangle, where x_e, x_d and x_n are the three vertices. Snyder designed his solvent scheme specially for the selection of mobile phases in chromatography. This separation technique is focused on the sequential optimization of both the total elution time of a multi-compound sample and the separation of the compounds within this sample. Therefore, Snyder suggested a solvent scheme where the strength of the solvent, which is related to the total elution time, and the selectivity of the solvent, which is associated with the kind of solute–

Table 3
Snyder's parameters for the selected solvent set

	Solvent	$P'x_e$	$P'x_d$	$P'x_n$
1	Diisopropyl ether	1.188	0.242	0.770
2	Di-*n*-butyl ether	0.901	0.136	0.663
3	Diethyl ether	1.595	0.319	0.986
4	Dioxane	1.824	1.008	1.968
5	Tetrahydrofuran	1.722	0.798	1.680
6	Anisole	0.980	1.085	1.435
7	Dibenzyl ether	0.891	0.891	1.518
8	Diphenyl ether	0.700	0.924	1.176
9	Fenetole	0.783	0.841	1.276
10	2-Butanone	1.620	0.765	2.115
11	Acetone	1.944	1.296	2.160
12	Ethyl acetate	1.462	1.075	1.806
13	Ethyl benzoate	1.452	1.188	1.760
14	Propylene carbonate	1.860	1.680	2.460
15	Dimethylacetamide	2.709	1.260	2.331
16	Dimethylformamide	2.624	1.344	2.432
17	N-Methylpyrrolidone	2.665	1.365	1.820
18	Tetramethylurea	2.300	0.700	2.000
19	Triethylamine	1.098	0.126	0.576
20	Dimethylsulfoxide	2.275	1.755	2.470
21	Hexamethylphos-phorotriamide	2.920	2.044	2.336
22	Nitrobenzene	1.350	1.215	1.935
23	Benzonitrile	1.610	1.196	1.794
24	Acetonitrile	2.046	1.612	2.542
25	Pyridine	2.279	1.113	1.908
26	2,6-Dimethylpyridine	2.021	0.774	1.505
27	Quinoline	2.080	1.404	1.716
28	Toluene	0.736	0.552	1.012
29	Benzene	0.870	0.840	1.290
30	Chlorobenzene	0.648	0.918	1.134
31	Bromobenzene	0.648	0.918	1.134
32	Carbon tetrachloride	0.510	0.646	0.544
33	1,2-Dichloroethane	1.332	0.703	1.665
34	Methylene chloride	1.156	0.578	1.666
35	Chloroform	1.232	1.716	1.452
36	*tert*-Butanol	2.145	0.897	0.858
37	Isopropanol	2.322	0.860	1.118
38	*n*-Butanol	2.067	0.819	1.014
39	Ethanol	2.652	1.092	1.456
40	Methanol	3.366	1.254	1.980
41	Ethylene glycol	2.538	1.242	1.620
42	Water	3.600	3.060	2.340

solvent interactions responsible for the separation between chromatographic peaks, could be adjusted separately. Thus, the P' parameter accounts for the solvent strength and x_e, x_d and x_n all describe the solvent selectivity.

The solvent-selectivity triangle and the proposed solvatochromic solvent scheme, though apparently analogous, do not allow a direct comparison. Whereas the former works with relative measures of solute–solvent interactions (i.e. x_e, x_d and x_n), the latter uses absolute parameters (i.e. α, β and π*). Therefore, a transformation must be applied either to obtain relative solvatochromic contributions or to have absolute Snyder parameters. The first possibility was chosen in a sound study reported by Snyder, Carr and Rutan [32]. In this work, the parameter related to the global interaction ability was defined as $\Sigma = \alpha+\beta+\pi^*$, and the relative solvatochromic contributions, also drawn in a planar triangular structure, as the normalized ratios α/Σ, β/Σ and π^*/Σ. Although globally rather similar, the solvatochromic triangle provides a better description of the solvent selectivity than that in Snyder's approach [31].

The separation between solvent strength and solvent selectivity, although very valuable in the chromatographic field, is not always the best approach for solvent selection. Indeed, in many other solvent-dependent processes (e.g., reaction kinetics or equilibria) both the strength and the nature of the solute–solvent interactions must be considered together to permit an understanding of the overall solvent effect on the process studied. In these situations, the use of a solvent classification based on absolute parameters appears to be the most suitable option. A comparison between the (absolute) solvatochromic scheme presented here and the Snyder classification requires the use of the original distribution constants (i.e. $P'x_e$, $P'x_d$ and $P'x_n$) related to the latter approach. In Table 3 are collected the $P'x_i$ values for the 42 solvents studied. The three-dimensional plot of this new solvent space ($P'x_e$, $P'x_d$ and $P'x_n$ now being the X, Y and Z axes) shows a continuous distribution (Fig. 5). At first sight, it seems dangerous to establish solvent classes, since there are no noticeable groups in the original solvent space and the clusters that could arise are likely to be mathematical artefacts. Moreover, the clear correlation which exists among the $P'x_i$ parameters does not suggest their use as clustering variables in clustering procedures where a Euclidean distance is taken as a similarity coefficient. The scores plot coming from principal component analysis could not reveal any kind of hidden class structure in the original data, either.

The reason behind the mathematical correlation of the Snyder parameters lies in the lack of independence among their chemical meanings, which is directly related to the mixed nature of Snyder's reference solutes. Thus, ethanol, dioxane and nitromethane were supposed to be sensitive only to the solvent hydrogen-bond basicity, the solvent hydrogen-bond acidity and the solvent polarity interactions, respectively, and this is far from being true.

The correlation analysis between the $P'x_i$ variables and the solvatochromic parameters shows no direct one-to-one correspondence (see Fig. 1). As could be expected, $P'x_d$ correlates with α and π^*, since dioxane can develop hydrogen-bond basic interactions (related to the solvent hydrogen-bond acidity, α) and polar interactions, whereas ethanol correlates with α, β and π^*, because it shows polar interactions and both hydrogen-bond interactions owing to its amphiprotic nature. When a combined treatment of stepwise and robust regression is performed [33] each one of the $P'x_i$ variables is modelled as follows:

$P'x_n=(2.27 \pm 0.08)\pi^*$
$r^2=0.96$ Scale estimate = 0.35
% points included in the model: 97.6%*

$P'x_d=(0.3 \pm 0.1)\alpha+(1.34 \pm 0.06)\pi^*$
$r^2=0.95$ Scale estimate = 0.23
% points included in the model: 92.8%*

$P'x_e=(1.8 \pm 0.2)\alpha+(0.8 \pm 0.1)\pi^*+(2.1 \pm 0.2)\beta$
$r^2=0.98$ Scale estimate = 0.28
% points included in the model: 85.7%*

*In the three equations, this percentage excludes the outlier observations.

The confirmed merged nature of $P'x_d$ and $P'x_e$ makes it difficult to interpret which concrete interaction causes lower or higher values of these parameters for each solvent.

The solvatochromic and Snyder approaches share the same underlying philosophy, i.e. they have the aim of classifying solvents in a microscopic way, by looking at their ability to develop the most essential interactions (hydrogen-bond acidity, hydrogen-bond basicity and polarity) with solutes. Although they are fairly similar for selectivity-based purposes, the solvatochromic approach is clearly more powerful for dealing with absolute solute–solvent interactions. The mixed nature of the Snyder parameters is most probably the cause of the loss of structure in the original solvent space, the swarm of points obtained being unsuitable for establishing a reliable structure of solvent classes.

5. Conclusions

The proposed solvent classification has been established by taking into account the solvent properties in the solvation sphere of solutes. Since these properties are often sufficient to describe the solvent effects in many chemical processes by means of LFER, the scheme is presented as a potential tool to be applied in the essential task of solvent selection. The absolute nature of the solvent descriptors makes the scheme suitable for a large number of chemical processes and the clear definition of the solvent classes facilitates the inclusion of new solvents in the groups previously proposed.

Acknowledgements

This work has received financial support from the Spanish Government through the DGICYT PB93-0744 and DGICYT PB93-1055 projects.

References

[1] C. Reichardt, Solvents and Solvent Effects in Organic Chemistry, VCH, Weinheim, 2nd edn., 1988.
[2] E.M. Kosower. J. Am. Chem. Soc., 80 (1958) 3253.
[3] K. Dimroth, C. Reichardt, T. Siepmann and F. Bohlmann, Justus Liebigs Ann. Chem., 661 (1963) 1.
[4] R.W. Taft and M.J. Kamlet, J. Am. Chem. Soc., 98 (1976) 2886.
[5] M.J. Kamlet and R.W. Taft, J. Am. Chem. Soc., 98 (1976) 377.
[6] M.J. Kamlet, J.L. Abboud and R.W. Taft, J. Am. Chem. Soc., 99 (1977) 6027.
[7] V. Gutmann, The Donor–Acceptor Approach to Molecular Interactions, Plenum Press, New York, 1978.
[8] C. Hansch and A. Leo, Substituent Constants for Correlation Analyses in Chemistry and Biology, Wiley-Interscience, New York, 1979.
[9] E. Grunwald and S.J. Winstein, J. Am. Chem. Soc., 70 (1948) 846.
[10] I.A. Koppel and A. Paju, Org. React. (Tartu), 11 (1974) 137.
[11] M. Krygowski and W.R. Fawcett, J. Am. Chem. Soc., 97 (1975) 2143.
[12] J.L.M. Abboud, R.W. Taft and M.J. Kamlet, J. Chem. Soc. Perkin Trans. 2, (1985) 15.
[13] W.J. Cheong and P.W. Carr, Anal. Chem., 60 (1988) 820.
[14] E. Casassas, G. Fonrodona and A. de Juan, An. Quim., 87 (1991) 611.
[15] D.L. Massart and L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.

[16] M. Chastrette, M. Rajzmann, M. Chanon and K.F. Purcell, J. Am. Chem. Soc., 107 (1985) 1.

[17] O. Pytela, Collect. Czech. Chem. Commun., 55 (1990) 644.

[18] M.J. Kamlet and R.W. Taft, Acta Chem. Scand. Ser. B, 39 (1985) 611.

[19] E. Casassas, G. Fonrodona, A. de Juan and R. Tauler, Chemom. Intell. Lab. Syst., 12 (1991) 29.

[20] M.H. Abraham, Pure Appl. Chem., 65 (1993) 2503.

[21] L.C. Tan and P.W. Carr, J. Chromatogr. A, 656 (1993) 521

[22] P. Meyer and G. Maurer, Ind. Eng. Chem. Res., 34 (1995) 373.

[23] M. Forina, R. Leardi, C. Armanino and S. Lanteri, PARVUS, an Extendable Package of Programs for Data Exploration, Classification and Correlation, Elsevier, Amsterdam, 1988.

[24] P.H.A. Sneath, J. Gen. Microbiol., 17 (1957) 201.

[25] P.H.A. Sneath and R.R. Sokal, Numerical Taxonomy: The Principles and Practice of Numerical Classification, Freeman, San Francisco, CA, 1973.

[26] G.N. Lance and W.T. Williams, Aust. Comput. J., 1 (1967) 15.

[27] J.H. Ward, J. Am. Stat. Assoc., 58 (1963) 236.

[28] D. Wishart, An Improved Multivariate Mode-Seeking Cluster Method, RSS Conference on Multivariate Analysis and its Applications, Hull, 1973.

[29] D. Wishart, in A.J. Cole (Editor), Mode Analysis in Numerical Taxonomy, Academic Press, New York, 1969.

[30] D. Wishart, Clustan User Manual, University of St. Andrews, Edinburgh, 4th edn., 1987.

[31] L.R. Snyder, J. Chromatogr., 92 (1974) 223.

[32] L.R. Snyder, P.W. Carr and S.C. Rutan. J. Chromatogr. A, 656 (1993) 537.

[33] E. Casassas, N. Domínguez, G. Fonrodona and A. de Juan, Anal. Chim. Acta, 283 (1993) 548.

The authors are at the Departament de Química Analítica, Universitat de Barcelona, Diagonal 647, 08028 Barcelona, Spain.

CHAPTER 5.

GOING FROM PURE SOLVENTS TO SOLVENT MIXTURES: THE WATER-DIOXANE EXAMPLE

5.1. Microscopic characterization of the mixture and determination of acid-base equilibria.

Correlation of acid–base properties of solutes with the polarity parameters and other solvatochromic parameters of dioxane–water mixtures

E. Casassas, G. Fonrodona and A. de Juan

Departament de Química Analítica, Universitat de Barcelona, Diagonal 647, 08028 Barcelona (Spain)

(Received January 30, 1991; revised May 8, 1991)

Abstract

The behaviour of water–dioxane mixtures as solvents is studied paying special attention to the correlations which may exist among the solvatochromic parameters of these mixtures, as well as to the correlations among some of these parameters and the acid–base properties of model solutes (whose acidity constants are determined). Values of the polarity parameter π^*, proposed by Kamlet and Taft, are determined from measurements of shifts in λ_{max} of 2-nitroanisole, 4-nitroanisole and 4-ethylnitrobenzene, proposed as reference dyes for amphiprotic solvents. The correlations of π^* values with some bulk properties and different microscopic parameters of the working solvents are established. Acid dissociation constants for propionic acid, chosen as a model for aliphatic carboxylic acids, and salicylic acid, for both aromatic carboxylic and phenolic hydroxylic acids, in water–dioxane mixtures are determined from e.m.f. data (at a constant electrolyte concentration 0.2 M and at 25 °C). The solvent composition range studied comprises from 10 to 70% dioxane (vol./vol.). Correlations obtained among the values of acid dissociation constants and the solvatochromic parameters for the solvent mixtures can be useful in helping to explain the variation of the acid–base character of solutes when properties of solvent are modified.

Introduction

Solvent mixtures have become an important subject of research because of their frequent use and the wide field of applications they offer. The most important feature of these mixed solvents is the gradual variation of properties they show when their composition is gradually modified.

A very interesting binary mixture is the water–dioxane mixture, first introduced in solution chemistry by Calvin and Wilson [1] and later used and studied by many authors [2, 3]. The difference between both constituents, especially with reference to their relative permittivity ($\epsilon = 78.54$ for pure water and 2.2 for pure dioxane), gives to their mixtures an unusually big span of properties. Thus, many different applications of this solvent mixture exist because of its versatility, among them, the study of the influence of decreasing polarity on certain phenomena related with biologic macromolecules [4] (stacking, hydrophobic interactions,...) or its use as a solvent medium to reach the differentiating titration of various functional groups in certain natural polyelectrolytes (humic and fulvic acids). Many studies have been made to determine macroscopic parameters (relative permittivity [5, 6], refraction index,...) or to try to determine the bulk structure of these hydroorganic mixtures, but a lack of investigation of the microscopic characteristics of the cybotactic zone of solutes in these solvents is noticed.

It is well-known fact that certain microscopic phenomena of solutes, such as proton-transfer equilibria, are strongly influenced by the solvent which forms their solvation sphere, which normally does not have the same properties of the bulk solvent nor frequently the same composition. In order to characterize this important zone of the solvent, a new group of microscopic parameters was proposed, the so-called solvatochromic parameters. Microscopic parameters can be determined from the values of many different kinds of properties (kinetic, thermodynamic,.....), but the easiest and most commonly used procedure involves the use of some reference dyes, called solvatochromic indicators, whose chromophores change the frequency of their absorption band when the properties of the solvent are modified. Among all the microscopic parameters, the most accepted and used are the following: $E_T(30)$, proposed by Dimroth and Reichardt [7], taken as a solvent polarity measure, and α, β and π^*, proposed by Kamlet and

II-12

Taft [8–10], to measure respectively the hydrogen-bond acidity, the hydrogen-bond basicity and the polarity of solvent.

In order to overcome the scarceness of information about the microscopic environment of the solutes in water–dioxane mixtures, the values of the solvatochromic parameters of these solvent systems are determined in the present work and correlations between all of them are proposed or confirmed, according to previous models [11].

Thus, in this work two different kinds of experiments have been carried out: the spectroscopic determination of the π^* parameter for water–dioxane mixtures, (α, β and $E_T(30)$ are already known from previous studies [12–14]) and the e.m.f. determination of the acidity constants of certain model solutes, taken as representatives of different functional groups. Between both sets of results interesting correlations have been obtained that can help to clarify the acid–base behaviour of the solutes in this hydroorganic mixture; behaviour that is clearly influenced by the properties of the cybotactic zone of the solutes.

Experimental

Reagents

Dioxane (PROBUS, a.r.) was purified by Eigenberger's method [15]. Water was deionized and distilled twice over potassium permanganate. 4-Nitroanisole (Merck, z.s.) was purified by active-carbon treatment in an acetone solution and recrystallized from water. 2-Nitroanisole, 4-ethylnitrobenzene and *N*-methyl-2-nitroaniline from Aldrich (a.r.). Potassium hydroxide (Merck, a.r.) CO_2-free solution in dioxane–water was prepared by the ion-exchange procedure [16]. Salicylic acid (Merck, a.r.) was purified by sublimation. Propionic acid and all other reagents were Merck, a.r.

Apparatus

A Beckman DU-7 spectrophotometer was interfaced (RS232) to an IBM personal computer. Spectra acquisition was controlled through Beckman data capture software. An ORION SA 720 potentiometer (precision ±0.1 mV) was used. An ORION 90-05 AgCl/Ag reference electrode with a ceramic junction and internal reference solution of sat. KCl in the working hydroorganic mixture was used in conjunction with an ORION 91-01 glass electrode. A double-walled cell was thermostatted at (25±0.1) °C. A Metrohm 665 Dosimat autoburette (precision 0.01 ml) with an exchange unit of 5 cm^3 was fitted with an antidiffusion burette tip. A magnetic stirrer was used. All of the titration apparatus was connected to a PC computer (HP Vectra ES/12 or HP 9133) through an interface HP 3421A, which allowed the full automatization of the titration process.

Procedure

Determination of the polarity–polarizability π^* parameter

For the determination of the polarity–polarizability π^* parameter of the studied water–dioxane mixtures, 2-nitroanisole (2-na), 4-nitroanisole (4-na), 4-ethylnitrobenzene (4-enb) and *N*-methyl-2-nitroaniline (Nm2na) were used as solvatochromic indicators. All of them were proposed by Kamlet *et al.* [10] for use in amphiprotic solvents or amphiprotic mixtures of solvents.

The spectrum of each test solution (a solution of a solvatochromic indicator in a given dioxane–water mixture) is recorded against a blank consisting of a dioxane–water mixture of identical composition as the solvent used in the test solution. Three replicates of each spectrum are obtained from identical independently prepared test solutions. From the digitalized average spectra, the wavelength of the longest-wavelength absorption maximum is determined. For every solvent composition this procedure is followed at three different concentration levels of the indicator. The gross average of the wavelengths of maximum absorption at each solvent composition is taken as the final value for the calculation of solvatochromic parameters.

From the gross average wavelength for each binary mixture, the values for the π^* parameter are obtained by the following expression

$$\pi^* = (\nu - \nu_0)/s$$

where ν is the frequency associated with the experimental wavelength, ν_0 is the frequency of the absorption maximum of the solvatochromic indicator dissolved in cyclohexane ($\pi^* = 0$) and s is the susceptibility of the measured property to changes in the solvent polarity.

Determination of acid dissociation constants in dioxane–water mixtures

Acid protonation constants for propionic acid and salicylic acid were determined from e.m.f. readings. The potentiometric cell used was: GE/working soln., 0.2 M, n% dioxane/RE (KCl_{sat}, n% dioxane) where GE is the glass electrode, RE the reference electrode and n is the dioxane percentage in volume in the working mixture. The solvent of the inner solution of the reference electrode has the same composition as that of the working mixture to minimize liquid junction potential problems which would originate from differences in solvents.

The Gran method [17] was used for *in situ* calibration of the cell and determination of the standard potential of the working electrode. The ionic product of the medium and the pH dependence of the liquid junction potential were also obtained from the strong acid–strong base titrations performed for calibration. For each mixture, the ionic product and the liquid junction potential were determined simultaneously through an iterative process using the MINIGLASS program [18]. Though liquid junction potentials (assumed to be given by expressions such as $E_J = j_H[H^+]$ or $E_J = j_{OH}[OH^-]$ in acid or basic media, respectively [19]) cannot be strictly considered constant with time for a ceramic junction, the small modifications in the value of the coefficient j_{OH} do not affect significantly the pK_a values calculated considering it as a constant, since the use of a ceramic junction minimizes its value and the period of time needed to perform the titrations is rather short. After calibration, a known amount of acid solute whose pK_a is to be determined was added to the cell and the titration continued until the suitable pH value was reached.

For all the deprotonation processes studied, several replicate experiments were performed. The experimental conditions of the titrations performed are listed in Table 1. As the deprotonation of both functional groups of salicylic acid occur in very different pH ranges, they can be studied in separated experiments, starting from salicylic acid or sodium salicylate solutions, depending on the studied process.

In all the titrations, the titrand and the titrant were prepared in solvent mixtures of the same composition and at the same total electrolyte concentration (0.2 M in KNO_3). The working solutions were titrated at 25 °C under a continuous flow of nitrogen.

In order to determine the pK_a values, the numerical analysis of e.m.f. data was carried out using the SUPERQUAD program [20]. The accuracy of the results is indicated by the parameter σ, which represents the ratio of the root mean square of the weighted residuals to the estimated error in the working conditions (0.01 ml for the autoburette volume readings and 0.1 mV for the e.m.f. readings),

TABLE 1. Experimental conditions of potentiometric titrations

Dioxane (%)	System	No. of titrations	Ligand concentration or concentration range (mmol l^{-1})	Range of $-\log[H^+]$
10	Prop-H[a]	4	6.9	3.0–6.5
	Sal$_1$-H[b]	3	5.6	2.5–5.5
	Sal$_2$-H[c]	4	24.8–25.0	10.2–12.5
20	Prop-H	3	6.9	3.5–7.0
	Sal$_1$-H	4	8.0	2.5–6.0
	Sal$_2$-H	4	12.5	10.0–12.6
30	Prop-H	3	34.5	3.5–7.0
	Sal$_1$-H	4	27.8–28.1	2.5–6.0
	Sal$_2$-H	3	12.3	11.0–13.0
40	Prop-H	3	32.8	3.5–7.5
	Sal$_1$-H	3	26.0–26.2	2.5–6.0
	Sal$_2$-H	3	17.9–18.4	11.5–13.3
50	Prop-H	3	32.4–32.6	4.0–8.0
	Sal$_1$-H	3	8.0	3.0–6.0
	Sal$_2$-H	3	17.9–18.0	11.7–13.7
60	Prop-H	3	32.2–32.8	4.0–8.0
	Sal$_1$-H	3	26.1	3.0–6.0
	Sal$_2$-H	3	18.0	12.0–14.0
65	Prop-H	3	34.5	4.5–8.0
	Sal$_1$-H	4	5.6–27.9	3.0–6.3
	Sal$_2$-H	3	17.6–17.8	12.2–14.5
70	Prop-H	3	34.5	5.0–9.0
	Sal$_1$-H	4	27.8	3.0–7.0
	Sal$_2$-H	3	18.0–18.1	12.5–14.7

[a] Propionic-H. [b] Salicylic-H (– COOH group). [c] Salicylic-H (– OH group).

and by the value of the statistic parameter χ^2, which is based on weighted residuals of e.m.f. readings [20].

Liquid junction potential, given by $E_J = j_{OH}[OH^-]$, must be taken into account in the calculation only for the deprotonation of the phenolic group of salicylic acid, which occurs in a very basic pH range. In the pH ranges where the deprotonation of carboxylic groups of both propionic and salicylic acids occurs, the influence of the liquid junction potential can be considered negligible.

Results

Table 2 shows the π^* values obtained from each solvatochromic indicator for each composition mixture (average of values obtained at three different concentration levels), and the gross-averaged π^* values, calculated as a mean from the three valid reference dyes, after rejecting the significantly different results of N-methyl-2-nitroaniline.

Table 3 gives the pK_w and j_{OH} values for the different water–dioxane mixtures, at 25 °C and a constant electrolyte concentration 0.2 M, evaluated using the MINIGLASS program.

Table 4 includes the pK_a values for salicylic and propionic acids for the different water–dioxane mixtures, at 25 °C and a constant electrolyte concentration 0.2 M, evaluated using the SUPERQUAD program.

Discussion

It can be observed in Table 2 that *N*-methyl-2-nitroaniline yields results which are very different from those obtained with the three other indicators. The differences are significant, as proved by the Cheong and Carr graphic method [11, 21]. From the results obtained in the factor analysis of all the data [22], it must be accepted that these differences arise from the effect of hydrogen-bond interactions with the solvent on the spectral shift of this indicator. Therefore, the π^* values obtained with this dye were rejected.

Although the results obtained from each one of the three valid reference dyes show small differences from each other, attributable mainly to differences in their basic character, which is of course very weak in all cases, other parameters proposed as a polarity measure are affected to a much more significative degree by the hydrogen-bond interactions. This is the case of $E_T(30)$, proposed by Dimroth and Reichardt, evaluated from 2,6-diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenoxide taken as a solvatochromic indicator. The structure of this substance favours the hydrogen-bond interactions between its phenoxide group and hydrogen atoms of the solvent. Thus, $E_T(30)$ is considered to be a mixed measure depending on the polarity of the solvent and also on its hydrogen-bond donor ability. This was observed by Kamlet and co-workers, who proved that $E_T(30)$ is a linear combination of π^* and α [23], which are the parameters proposed by them as a measure of

TABLE 2. π^* values obtained from each solvatochromic indicator tested and averaged π^* values for the different compositions of water–dioxane mixtures

Composition of the solvent mixture		π^*				π^*_{average}
% dioxane	n_2	4-na	2-na	4-enb	Nm2na[a]	
100	1.000	0.537	0.541	0.543	0.585	0.540
95	0.801	0.637	0.642	0.631	0.706	0.636
90	0.655	0.718	0.714	0.680	0.771	0.704
85	0.545	0.762	0.765	0.718	0.820	0.748
80	0.458	0.787	0.802	0.755	0.870	0.781
75	0.388	0.840	0.836	0.795	0.910	0.823
70	0.330	0.857	0.874	0.815	0.953	0.849
65	0.282	0.890	0.902	0.863	1.017	0.885
60	0.241	0.924	0.935	0.902	1.025	0.920
50	0.174	0.974	1.015	0.978	1.134	0.989
40	0.123	1.042	1.065	1.039	1.227	1.049
30	0.083	1.071	1.113	1.080	1.293	1.088
20	0.050	1.089	1.152	1.132	1.353	1.124
10	0.023	1.106	1.169	1.154	1.366	1.143
5	0.011	1.111	1.177	1.159	1.375	1.149

[a]These values are not included in the final averaged values.

TABLE 3. Ionic product, pK_w, and liquid junction potential coefficient, j_{OH}, for dioxane–water mixtures of 25 °C and a constant electrolyte concentration of 0.2 M, calculated using MINIGLASS program

Dioxane (%)	pK_w	Standard deviation	j_{OH}[a]	Standard deviation
10	13.963	0.004	63	5
20	14.149	0.005	82	22
30	14.3880	0.0004	46	4
40	14.740	0.002	98	3
50	15.100	0.003	103	3
60	15.525	0.005	31	3
65	15.864	0.005	88	2
70	16.09	0.02	80	7

[a]Values of j_{OH} are given in mV l mol^{-1}.

the polarity and the hydrogen-bond acidity of solvent, respectively. The different nature of $E_T(30)$ and π^*, both proposed as polarity measures, can be seen easily from the plot of π^* versus $E_T(30)$ (Fig. 1), where the non-linearity between the homologous values of each parameter is evident.

In a recent investigation, Cheong and Carr [11] confirmed the mixed nature of $E_T(30)$ and proposed the following general equation, valid for several hydroorganic mixtures: $E_T(30)=31.92(\pm 3.8)+11.42(\pm 5.6)\pi^*+15.96(\pm 3.0)\alpha$ (in parentheses, the standard deviation associated to the coefficient).

The relationship obtained for water–dioxane mixtures from a multiple regression analysis using the π^* data obtained in this work agrees with the Cheong and Carr equation and can be expressed as follows:

$$E_T(30)=31.75+10.83\pi^*+18.06\alpha$$

$$\text{standard deviation}=0.295 \qquad r=0.9992$$

As a consequence of the mixed nature of $E_T(30)$, which is sensitive to many solute–solvent interactions, this parameter has been used to investigate the phenomena of preferential solvation in many binary mixtures [24, 25].

According to Dawber *et al.* [25], if the components of a binary solvent mixture participate randomly in the solvation of the solute, a linear relationship analogous to the Raoult law is expected: $E_T(30)_{\text{mixture}}=\Sigma X_i E_T(30)_i^0$, where $E_T(30)_i^0$ is the value

TABLE 4. Values of logarithm of protonation constants for propionic and salicylic acids at 25 °C and a constant electrolyte concentration of 0.2 M, in water–dioxane mixtures, calculated using SUPERQUAD program

Dioxane (%)	System	log *K*	Standard deviation	σ	χ^2
10	Prop-H[a]	4.825	0.003	1.534	17.64
	Sal_1-H[b]	2.897	0.001	1.590	23.26
	Sal_2-H[c]	12.789	0.008	1.952	18.35
20	Prop-H	5.091	0.006	2.235	24.36
	Sal_1-H	3.067	0.003	1.482	7.07
	Sal_2-H	13.23	0.02	1.010	4.19
30	Prop-H	5.395	0.001	2.450	37.26
	Sal_1-H	3.318	0.002	2.863	49.95
	Sal_2-H	13.43	0.02	1.006	16.23
40	Prop-H	5.6762	0.0006	1.722	28.48
	Sal_1-H	3.524	0.003	2.354	17.33
	Sal_2-H	13.51	0.02	2.052	1.55
50	Prop-H	6.063	0.001	3.423	18.63
	Sal_1-H	3.789	0.006	2.736	5.67
	Sal_2-H	13.94	0.02	1.604	4.95
60	Prop-H	6.4940	0.0005	1.275	32.69
	Sal_1-H	4.305	0.001	2.318	18.69
	Sal_2-H	14.12	0.02	2.086	4.67
65	Prop-H	6.745	0.001	2.195	6.01
	Sal_1-H	4.478	0.003	2.359	56.41
	Sal_2-H	14.46	0.01	1.365	4.84
70	Prop-H	6.879	0.002	2.980	41.77
	Sal_1-H	4.654	0.001	2.830	17.48
	Sal_2-H	14.84	0.02	3.228	27.41

[a]Propionic-H. [b]Salicylic-H (–COOH group). [c]Salicylic-H (–OH group).

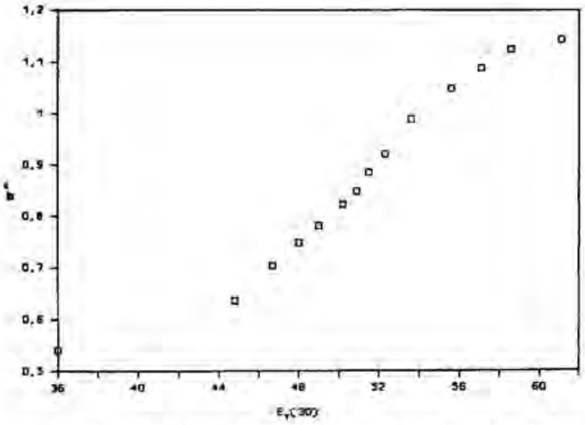


Fig. 1. Plot of π^* vs. $E_T(30)$ for water–dioxane mixtures. π^* values were taken from Table 1; $E_T(30)$ values are from the literature [12, 13].

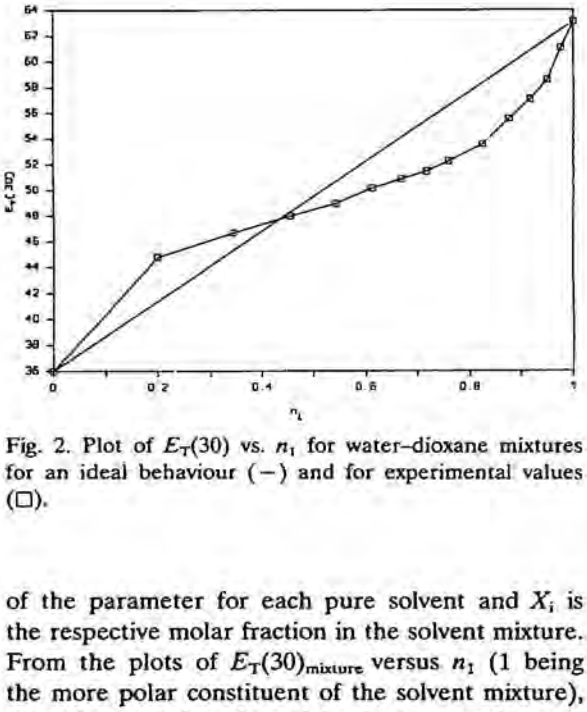


Fig. 2. Plot of $E_T(30)$ vs. n_1 for water–dioxane mixtures for an ideal behaviour (–) and for experimental values (□).

of the parameter for each pure solvent and X_i is the respective molar fraction in the solvent mixture. From the plots of $E_T(30)_{\text{mixture}}$ versus n_1 (1 being the more polar constituent of the solvent mixture), the existence of preferential solvation can be seen when a loss of linearity is produced ('preferential' meaning the presence in the cybotactic zone of a higher content of the constituent than that predicted by the linearity assumption).

In the case of water–dioxane mixtures (Fig. 2), a double behaviour is noticed. If this assumption is correct, for solutions with $n_2 < 0.55$ (% dioxane < 85), the preferential solvation is due to dioxane, and for higher percentages of this cosolvent, water is the constituent responsible of this phenomenon.

Knowledge of the different microscopic parameters of dioxane–water mixtures and the existence of correlations as above allow an approach to be made to the properties and structure of the cybotactic zone of solutes. It would be interesting to compare these characteristics with those present in the bulk of these solvent mixtures.

In the plot of $1/\epsilon$ versus π^* (Fig. 3), both taken as pure measures of polarity, the first one referring to the bulk of the solvent and the second one related with the cybotactic zone of the solute, a colinearity can be observed for $\epsilon > 23$ ($1/\epsilon < 0.037$), i.e. for percentages of dioxane (vol./vol.) < 60%, approximately. When values of ϵ decrease, that is, for higher percentages of dioxane, a clear deviation of the linearity can be detected. This means that in mixtures richer in dioxane (% dioxane > 60), the polarity varies in a different way than in mixtures poorer in this cosolvent. The behaviour shown by the Figure points to an increased polarity of the cybotactic zone (measured by π^*) in mixtures richer in dioxane in comparison with the predicted value given by the linear correlation found for the poorer mixtures. These different properties of the bulk solvent and the solvation sphere may be the reason for the inadequacy of certain theories that try to explain the proton-transfer microscopic process as a function only of macroscopic properties of the solvent mixture, as n_2 or $1/\epsilon$.

Thus, in the plot of pK_w values (Table 3) or pK_a values of propionic and salicylic acid (Table 4) versus molar fraction of dioxane (Fig. 4) or $1/\epsilon$ (Fig. 5) it can be seen that no linear correlations are obtained that are valid for the whole range of studied compositions. It is in the zone in which deviations start to be observed where the specific solute–solvent interactions become more important and the ineffectiveness of an only macroscopic model more evident.

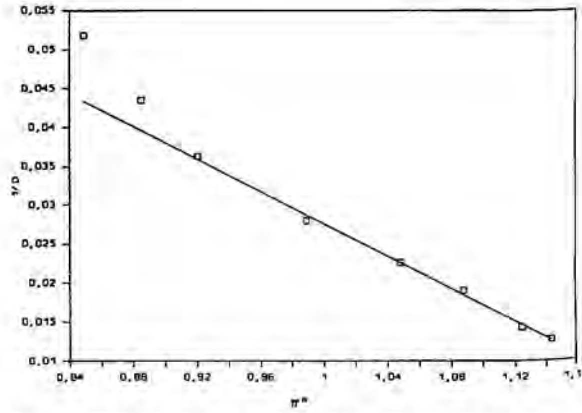


Fig. 3. Plot of $1/\epsilon$ vs. π^* for water–dioxane mixtures. π^* values are those from Table 1; ϵ values are taken from the literature [6].

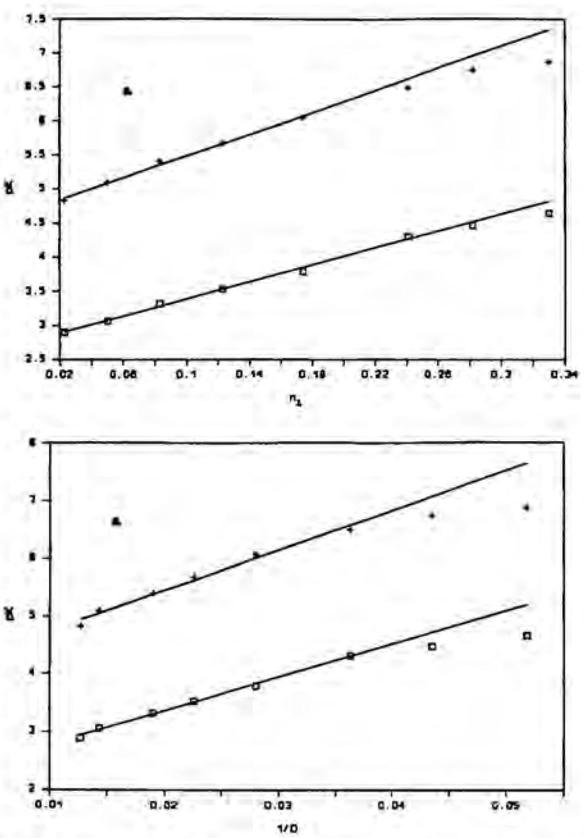


Fig. 4. Plot of pK_a values of (a) carboxylic groups of salicylic acid (□) and propionic acid (+); (b) phenolic group of salicylic acid (□) and pK_w values (+) vs. molar fraction of dioxane.

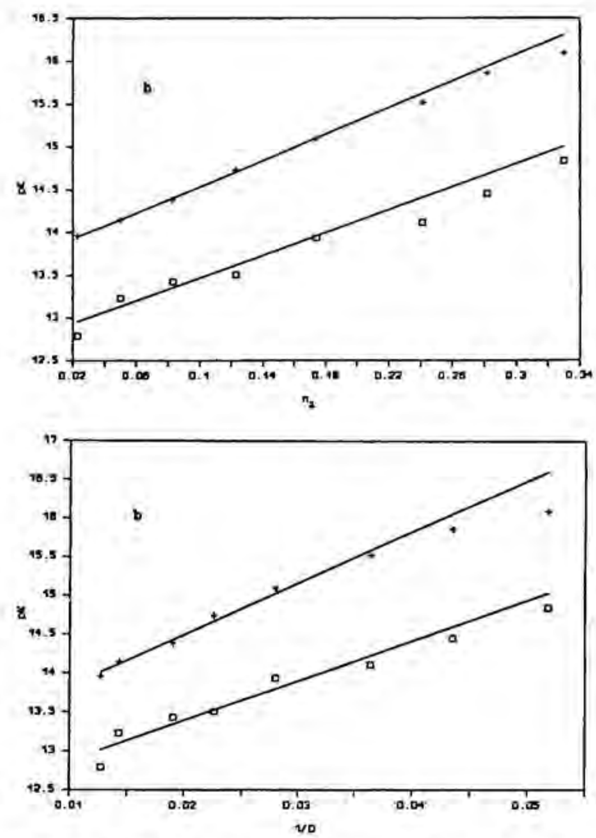


Fig. 5. Plot of pK_a values of (a) carboxylic groups of salicylic acid (□) and propionic acid (+); (b) phenolic group of salicylic acid (□) and pK_w values (+) vs. $1/\epsilon$ for water–dioxane mixtures.

So, research which could lead to expressions that relate the acidity constants of solutes with the properties of their immediate environment, well-characterized by the microscopic parameters, is justified.

When the plots of pK_w and pK_a values versus $E_T(30)$ (Fig. 6) are depicted two linear segments can be observed, whose intersection is between 50 and 60% of dioxane. The different behaviour of both zones can be attributed to the breakdown (detected by Langhals [12]) of the hydrogen-bridged structure of water by dioxane when the amount of this cosolvent begins to be high. Since it has been previously proved that $E_T(30)$ measures the polarity and the hydrogen-bond donor acidity of the solvent and since this last property will change strongly with the destruction of the water structure, it is clear that two differentiated trends will result in the behaviour of acid solutes, well noticeable in the last mentioned Figures.

A more general approach is that of Kamlet and Taft, who suggested a general equation [10] which explains any solute property varying with solvent composition as a linear combination of the microscopic parameters of the solvent responsible for the modification of the solute property studied. This general equation has also been applied to solvent mixtures and is expressed very commonly as follows [10]: $XYZ=(XYZ)_0+a\alpha+b\beta+s\pi^*$, where α, β and π^* are the microscopic parameters previously described, XYZ is the solute property, XYZ_0 the value of this property for the same solute in a hypothetical solvent for which $\alpha=\beta=\pi^*=0$, and a, b and c are numerical coefficients related with the susceptibility of the studied solute property to changes in α, β and π^*, respectively. This equation can include additional terms or some of its terms can become equal to zero depending on the property of the solute to be described [26].

In the present work, several attempts have been made to find the best form of the Kamlet and Taft equation to describe the variation of pK values (for both the autoprotolysis and the acid dissociation processes) in water–dioxane mixtures.

Multiple regression analysis has been applied to our data. All possible combinations of solvatochromic parameters have been checked. The best fit is obtained when only the π^* parameter is used, yielding

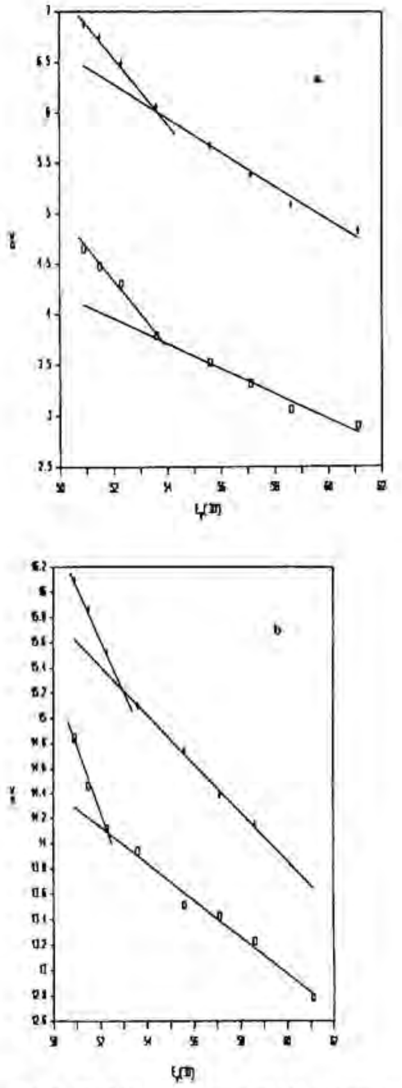


Fig. 6. Plot of pK_a values of (a) carboxylic groups of salicylic acid (□) and propionic acid (+); (b) phenolic group of salicylic acid (□) and pK_w values (+) vs. $E_T(30)$ for water–dioxane mixtures.

the following general equation

$$pK = s\pi^* + pK^0$$

For each one of the systems studied, this equation becomes

$$pK_{1sal} = -5.9(\pm 0.1)\pi^* + 9.68(\pm 0.04) \quad r = 0.998$$
$$pK_{2sal} = -5.9(\pm 0.5)\pi^* + 19.7(\pm 0.1) \quad r = 0.981$$
$$pK_{prop} = -6.8(\pm 0.3)\pi^* + 12.77(\pm 0.08) \quad r = 0.995$$
$$pK_w = -7.1(\pm 0.1)\pi^* + 22.10(\pm 0.04) \quad r = 0.998$$

(in parentheses, as before, the standard deviation associated with the coefficient).

These simple equations explain well all the experimental data within the composition range studied. If the three solvatochromic parameters are forced into the equation, no improvement is obtained in the fit. For instance, for the first acid dissociation of salicylic acid, the equation would be:

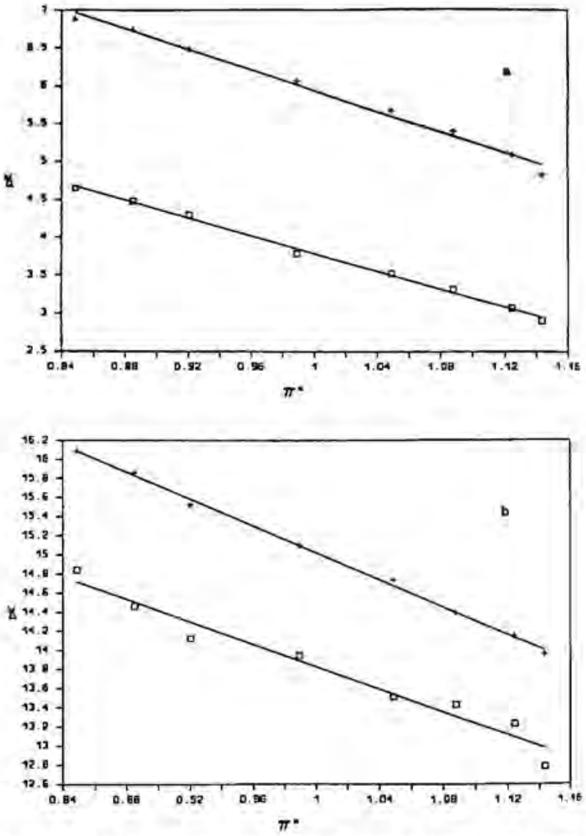


Fig. 7. Plot of pK_a values of (a) carboxylic groups of salicylic acid (□) and propionic acid (+); (b) phenolic group of salicylic acid (□) and pK_w values (+) vs. π^* for water–dioxane mixtures.

$$pK_{1sal} = -0.4(\pm 1)\alpha - 1(\pm 5)\beta - 6(\pm 2)\pi^* + 11.4$$

$$r = 0.997 \quad r_\alpha = 0.167$$
$$r_\beta = 0.155$$
$$r_\pi = 0.895$$

The small coefficients in the α and β terms and the big error associated with them, as well as the small partial correlation coefficients for these two parameters, confirm that their contribution in the description of the variation of pK values in water–dioxane mixtures can be neglected.

Thus, a really useful correlation is finally found, that shows the linear dependence of acid dissociation constants with the microscopic polarity of solvent (expressed by π^*) for the whole range of compositions in water–dioxane mixtures (Fig. 7). The previous correlation, which considered $1/\epsilon$ as a measure of polarity, has been modified by the introduction of the microscopic character of this solvent property, defined by π^* parameter, which so strongly affects the proton-transfer process.

Acknowledgements

This research was supported by a CICYT grant number PB87-0061 and by a CIRIT (Generalitat de Catalunya) grant.

References

1 M. Calvin and K. W. Wilson, *J. Am. Chem. Soc.,* (1945) 2003.

2 LeGrand G. Van Uitert and C. Haas, *J. Am. Chem. Soc., 75* (1953) 451, 455, 457, 2739.

3 S. C. Mohr, W. D. Wilk and G. H. Barrow, *J. Am. Chem. Soc., 87* (1965) 3048.

4 R. Tribolet, R. Malini-Balakrishnan and H. Sigel, *J. Chem. Soc., Dalton Trans.,* (1985) 2291.

5 G. Akerlöf and O. A. Short, *J. Am. Chem. Soc., 58* (1936) 1241.

6 F. E. Critchfield, J. A. Gibson and J. L. Hall, *J. Am. Chem. Soc., 58* (1953) 1241.

7 K. Dimroth and C. Reichardt, *Z. Anal. Chem., 215* (1966) 344.

8 R. W. Taft and M. J. Kamlet, *J. Am. chem. Soc., 98* (1976) 2886.

9 M. J. Kamlet and R. W. Taft, *J. Am. Chem. Soc., 98* (1976) 377.

10 M. J. Kamlet, J. L. Abboud and R. W. Taft, *J. Am. Chem. Soc., 99* (1977) 6027.

11 W. J. Cheong and P. W. Carr, *Anal. Chem., 60* (1988) 820.

12 H. Langhals, *Angew. Chem., Int. Ed. Engl., 21* (1982) 724.

13 E. Casassas and G. Fonrodona, *J. Chim. Phys., 86* (1989) 391.

14 E. Casassas, G. Fonrodona and A. de Juan, *J. Sol. Chem., 20* (12) (1991) in press.

15 A. I. Vogel, *A Textbook of Practical Organic Chemistry,* Longmans Green, London, 1989, 5th edn., p. 177.

16 J. E. Powell and M. A. Hiller, *J. Chem. Educ., 34* (1957) 330.

17 G. Gran, *Analyst, 77* (1952) 661.

18 J. L. Beltrán and A. Izquierdo, *Anal. Chim. Acta, 181* (1986) 87.

19 G. Biedermann and L. G. Sillén, *Ark. Kemi, 5* (1952) 425.

20 P. Gans, A. Sabatini and A. Vacca, *J. Chem. Soc., Dalton Trans.,* (1985) 1195.

21 E. Casassas, G. Fonrodona and A. de Juan, *An. Quim., 87* (1991) in press.

22 E. Casassas, G. Fonrodona, A. de Juan and R. Tauler, *Chemom. Intell. Lab. Syst.,* (1991) in press.

23 J. L. M. Abboud, R. W. Taft and M. J. Kamlet, *J. Chem. Soc., Perkin Trans. II,* (1985) 815.

24 J. R. Haak and J. B. F. N. Engberts, *Recl. Trav. Chim. Pays-Bas, 105* (1986) 307.

25 J. G. Dawber, J. Ward and R. A. Williams, *J. Chem. Soc., Faraday Trans. 1, 84* (1988) 713.

26 M. J. Kamlet and R. W. Taft, *Acta Chem. Scand., Ser. B, 39* (1985) 611.

DETERMINACION DEL PARAMETRO DE POLARIDAD-POLARIZABILIDAD π^* Y CORRELACION DE ESTE CON $E_T(30)$ PARA MEZCLAS DIOXANO-AGUA

POR

E. CASASSAS, G. FONRODONA y A. DE JUAN

Departament de Química Analítica. Universitat de Barcelona Diagonal 647. Barcelona 08028

Recibido el 17 de agosto de 1990

En homenaje al Dr. Alvaro Izquierdo Irazu con motivo de su jubilación.

RESUMEN.—En las mezclas dioxano-agua, cubriendo un intervalo de composiciones entre el 5 y el 100 % de dioxano (v/v), se ha procedido a la determinación del parámetro π^* de Kamlet y Taft, que mide la polaridad-polarizabilidad del disolvente. Para ello se han utilizado algunos indicadores solvatocrómicos elegidos entre los recomendados por Kamlet y Taft para disolventes anfipróticos, carácter otorgado por el agua a las mezclas estudiadas. También se ha estudiado la idoneidad de cada uno de los indicadores solvatocrómicos escogidos. Por último, se han comparado el parámetro π^* de Kamlet y Taft y el parámetro $E_T(30)$ de Reichardt, definidos como medidas de la polaridad del disolvente, confirmándose en este último la presencia de una contribución importante debida a interacciones de puente de hidrógeno por donación de protones, reflejada en la correlación existente entre ambos parámetros.

SUMMARY.—Values of the polarity parameter π^*, proposed by Kamlet and Taft, have been determined for dioxane-water mixtures from measurements of shifts in ν_{max} of referece dyes, so-called solvatochromic indicators, which are carefully chosen according to the properties of the solvent mixture studied, in order to avoid or minimize the effect of solvent-solute hydrogen bond interactions. Substances used for the water-dioxane mixtures have been 2-nitroanisol, 4-nitroanisol and 4-ethylnitrobenzene, all of them recommended by Kamlet and Taft for amphyprotic solvents. The mixture composition range covered goes from 5 to 100 % of dioxane (v/v). A correlation is obtained between π^* parameter and $E_T(30)$ parameter, both of them proposed as a measure of polarity. The expression obtained shows that $E_T(30)$ is not a pure measure of polarity, but includes a contribution related to hydrogen-bond donor acidity of solvent, α.

INTRODUCCION

Las características propias de las mezclas hidroorgánicas de disolventes, mezclas que suelen poseer un poder solubilizante de compuestos orgánicos mayor que el agua y que presentan una variación gradual de propiedades al modificar la proporción de sus componentes, han hecho del estudio de estas mezclas un importante campo de investigación en Química Analítica, orientado a una mayor caracterización de las mismas y al estudio de las aplicaciones que pueden derivar de su uso.

La creciente utilización de mezclas hidroorgánicas de disolventes ha requerido también un conocimiento más profundo de las interacciones que éstas ejercen sobre los solutos. Es bien sabido que el comportamiento de los compuestos que se hallan en disolución se ve afectado por el disolvente que compone la esfera de solvatación, el cual, a su vez, ve alterado el valor de sus propiedades con respecto al que éstas poseen en el seno de la disolución. Por ello, cada vez son más numerosos los intentos de determinar las características de los disolventes en la zona cibotáctica, características imposibles de cuantificar con los parámetros macroscópicos tradicionales, que sólo son reflejo de la estructura global de la disolución. La necesidad de superar esta limitación ha llevado a la introducción de otro tipo de parámetros, de carácter microscópico, cuya determinación se realiza a partir de la medida de ciertas propiedades de unos solutos de referencia que se ven modificadas al hacerlo la naturaleza o las características del disolvente que rodea cada una de las partículas (moléculas o iones) de soluto. Los datos experimentales más frecuentemente empleados con este fin proceden de transiciones espectrales de ciertos solutos de referencia, que reciben el nombre de indicadores solvatocrómicos. Se obtienen a partir de estos datos mediante el tratamiento matemático adecuado los llamados parámetros solvatocrómicos. Ha sido definida una gran variedad de parámetros de este tipo, entre los cuales los más aceptados y de mayor utilidad son el parámetro $E_T(30)$, propuesto por Dimroth y Reichardt como medida de la polaridad (1), y los parámetros α, β y π^* de Kamlet y Taft (2, 3, 4) que cuantifican la capacidad formadora de puentes de hidrógeno por donación (α) o aceptación de protones (β) y la polaridad-polarizabilidad (π^*).

La importante variación de la polaridad con la composición en el sistema binario dioxano-agua, constatada por el amplio intervalo de valores de la constante dieléctrica (global) que presenta este disolvente mixto, intervalo que va desde 2,2 para el dioxano puro hasta 74,58 para el agua pura, hace necesario un estudio más profundo de aquella variación de polaridad desde el punto de vista microscópico. Este estudio se realiza en el presente trabajo, donde se procede a la determinación del parámetro π^* para las diferentes mezclas binarias dioxano-agua, cubriendo un intervalo de composición del 5 al 100 % (v/v) de dioxano, e investigando la posible correlación existente entre estos valores y los del parámetro $E_T(30)$. Estos últimos, que son una medida de la energía de la transición electrónica del compuesto betaínico [2,6-difenil-4(2,4,6-trifenil)piridin-1-il]fenóxido disuelto en la mezcla binaria de trabajo, se han tomado de la bibliografía (5, 6).

El parámetro π^* (4) recibe esta designación por estar basado en la determinación de los desplazamientos que sufren ciertas transiciones espectrales en el visible o el ultravioleta en las que intervienen electrones π en niveles energéticos antienlazantes, $p \rightarrow \pi^*$ o $\pi \rightarrow \pi^*$.

Para que una substancia sea útil como indicador solvatocrómico debe cumplir una serie de características generales, que incluyen: presentar absorbancia en zonas accesibles del espectro, sufrir un desplazamiento análogo al del resto de indicadores en disolventes no formadores de puente de hidrógeno por donación de protones, poseer una frecuencia de absorción que no varíe por superposición con otras bandas de energía o por cambio de forma de la banda, y responder con suficiente sensibilidad a los cambios de polaridad del disolvente.

De todos modos, no todos los solutos que cumplan las condiciones anteriores pueden utilizarse para determinar el parámetro π^* de cualquier tipo de disolvente. Los indicadores idóneos serán aquellos cuya frecuencia de máxima absorción sólo se vea afectada por la variación de la propiedad del disolvente que se quiere cuantificar, en este caso, la polaridad. Esta aserción confirma la importancia de una correcta elección de los solutos de referencia, que debe conseguir eliminar todas las contribuciones ajenas a la propiedad en estudio, especialmente las procedentes de la posibilidad de formación de puentes de hidrógeno entre el indicador y el disolvente o mezcla de disolventes. Así pues, para el caso de mezclas binarias dioxano-agua, de carácter global anfiprótico conferido por la presencia de agua, los solutos de referencia a utilizar, por su propia estructura, han de ser incapaces de formar puentes de hidrógeno por donación o por aceptación de protones. De todas formas, aun cuando es sencillo fijar las condiciones teóricas necesarias para que un indicador sea correcto, no resulta tan fácil encontrar substancias que cumplan los requisitos deseados. Es especialmente difícil hallar compuestos útiles para disolventes anfipróticos, pues si bien existen muchos solutos incapaces de formar puentes de hidrógeno, son pocos los que aúnan esta característica a una sensibilidad suficientemente grande frente a los cambios de polaridad del disolvente. En este caso se adopta una solución de compromiso, combinando el uso de indicadores no formadores de puentes de hidrógeno con el uso de otras substancias de tan baja basicidad que permiten considerar nula su capacidad de formación de puentes de hidrógeno sin cometer errores significativos. Para el caso de la mezcla dioxano-agua hay que utilizar los indicadores recomendado por Kamlet y Taft para disolventes anfipróticos, que son: 2-nitroanisol (2-na), 4-nitroanisol (4-na), 4-etilnitrobenceno (4-enb) y N-metil-2-nitroanilina (nm2na).

PARTE EXPERIMENTAL

Reactivos y Soluciones

a) Disolventes.

Dioxano (PROBUS p.a) previamente purificado según el método de Eigenberger (7). Agua, previamente desionizada y bidestilada sobre permanganato.

b) Indicadores solvatocrómicos.

2-Nitroanisol (Aldrich, p.a.); 4-etilnitrobenceno (Aldrich, p.a.); N-metil-2-nitroanilina (Aldrich, p.a); 4-nitroanisol (Merck, p. s) purificado por reprecitación con agua de una solución en acetona, previamente tratada con carbón activo.

Las soluciones stock de los indicadores solvatocrómicos se preparan en dioxano. Las soluciones de trabajo se preparan a un mínimo de tres niveles de concentración diferentes (entre $5/10^{-4}$ y 10^{-5}M) a partir de las soluciones stock por dilución con la cantidad adecuada de dioxano y de agua para obtener la composición requerida del disolvente binario.

Aparatos

Espectrofotómetro BECMAN-DU equipado con cubetas de cuarzo de 1 cm de camino óptico y una rendija de 0,5 nm, conectado a un ordenador IBM-PC a través de una interfase RS232. La adquisición de los espectros se ha controlado utilizando el software Data Capture de Beckman.

Técnica de trabajo

El parámetro π^* se determina a partir de los espectros de los indicadores solvatocrómicos para cada composición de la mezcla binaria. Del espectro de la solución de indicador solvatocrómico corregido por el de un blanco que contiene el disolvente puro (una mezcla agua-dioxano a la misma composición que la empleada en la solución del indicador) se determina el máximo de absorción mediante los programas usuales de digitalización de espectros.

Con cada soluto de referencia se han llevado a cabo un mínimo de tres series independientes de espectros a diferente concentración de indicador para cada composición de la mezcla, número incrementado cuando los desplazamientos espectrales no eran suficientemente grandes o reproducibles.

En todos los casos se realiza un barrido de longitudes de onda que cubre un mínimo de 100 nm a una velocidad de 120 nm/min. Cada espectro contiene un total de 1.000 o más datos, ya que las lecturas de absorción se realizan a intervalos de 0,1 nm. Para el 2-nitroanisol, el 4-nitroanisol y el 4-etilnitrobenceno, el barrido de longitudes se ha realizado entre 250 y 350 nm y para la N-metil-2-nitroanilina entre 350 y 500 nm.

TABLA I

Valores experimentales de las longitudes de onda de máxima absorción para las series de espectros independientes realizadas con cada indicador (en el encabezamiento de cada columna se da la concentración de indicador, expresada en mol · l^{-1})

	1-etil-4-nitrobenceno λ_{max} (nm)			2-nitroanisol λ_{max} (nm)			N-metil-2-nitroanilina λ_{max} (nm)	
% dioxano	7,3 · 10^{-5}	5,8 · 10^{-5}	4,4 · 10^{-5}	8,1 · 10^{-5}	1,1 · 10^{-4}	1,6 · 10^{-4}	4,2 · 10^{-5}	2,1 · 10^{-4}
100	274,2	274,6	273,9	319,9	320,3	319,0	422,7	422,7
95	275,9	275,9	275,5	322,8	322,4	322,5	426,6	425,7
90	276,9	276,6	276,4	324,7	324,4	324,1	427,9	428,2
85	277,6	277,2	277,2	325,5	326,1	325,5	429,6	429,4
80	278,2	277,9	279,3	326,7	326,7	326,6	431,1	430,8
75	279,0	278,5	280,1	327,2	327,8	327,6	433,1	432,1
70	279,5	278,7	279,2	328,6	328,6	328,5	433,4	433,5
65	279,9	280,0	281,2	328,9	329,6	329,3	435,0	435,7
60	280,8	280,5	281,7	329,8	330,4	330,2	435,1	436,1
50	282,1	281,9	282,7	331,9	332,1	332,8	438,5	439,3
40	283,2	283,0	283,3	333,9	333,4	333,6	442,3	441,3
30	283,7	284,1	284,1	334,7	335,1	335,0	444,0	443,7
20	284,8	284,8	285,2	336,1	335,9	336,0	445,8	445,7
10	285,1	285,3	285,3	336,4	336,5	336,5	446,1	446,2
5	285,3	285,3	285,3	336,7	336,7	336,6	446,4	446,5

			4-nitroamisol λ_{max}(nm)			
% dioxano	3,04 · 10^{-5}	8,89 · 10^{-5}	3,04 · 10^{-5}	6,32 · 10^{-5}	8,89 · 10^{-5}	4,56 · 10^{-5}
100	304,5	304,0	304,5	304,3	304,2	304,3
95	306,5	306,5	306,5	306,5	306,4	306,5
90	308,5	308,5	308,5	308,0	307,9	308,3
85	309,5	310,0	309,5	308,4	308,9	309,3
80	309,5	310,5	309,5	309,7	309,9	309,8
75	310,7	312,0	311,2	310,7	310,5	311,0
70	311,0	312,0	311,0	311,2	311,8	311,4
65	312,0	312,5	311,7	312,0	312,6	312,2
60	312,5	312,5	312,5	313,8	313,4	312,9
50	313,5	314,0	313,5	314,3	315,2	314,1
40	314,5	315,5	316,0	316,0	316,3	315,7
30	315,5	315,5	316,0	317,9	316,8	316,3
20	316,0	316,5	316,0	318,2	317,1	316,8
10	316,5	316,5	317,0	318,5	317,4	317,2
5	316,5	316,8	316,7	318,6	317,9	317,3

Tratamiento numérico de los datos

El cálculo del parámetro π^* se realiza utilizando la expresión:

$$\pi^* = (\nu - \mu_o/s) \qquad [1]$$

donde ν^o es la frecuencia del máximo de absorción del indicador disuelto en ciclohexano, s es el factor de Kamlet y Taft que expresa la susceptibilidad de la propiedad medida frente a variaciones de la polaridad-polarizabilidad (ambos tabulados por Kamlet y Taft (4)) y ν es la frecuencia del máximo del indicador disuelto en el disolvente problema. Los valores de π^* para cada indicador individual se obtienen del valor medio de los resultados que provienen de sus diferentes series a cada composición de la mezcla binaria. El promedio global de los diversos valores de π^* procedentes de los indicadores empleados proporcionará los valores de dicho parámetro que serán adoptados como definitivos para posteriores correlaciones.

RESULTADOS Y DISCUSION

En la Tabla I se muestran las longitudes de onda de absorción máxima para cada uno de los indicadores, a las diferentes concentraciones empleadas y para cada composición del disolvente en estudio.

La Tabla II contiene los valores de π^* obtenidos a partir de los cuatro indicadores diferentes para

TABLA II

Valores del parámetro π^* para cada indicador y valor promedio de π^* para las diferentes mezclas de agua-dioxano.

Composición de la mezcla disolvente		4-na	2-na	4-enb	Nm2na	
% dioxano	n_2	π^*	π^*	π^*	$\pi^{*(*)}$	π^*
100	1.000	0.537	0.541	0.543	0.585	0.540
95	0.801	0.637	0.642	0.631	0.706	0.636
90	0.655	0.718	0.714	0.680	0.771	0.704
85	0.545	0.762	0.765	0.718	0.820	0.748
80	0.458	0.787	0.802	0.755	0.870	0.781
75	0.388	0.840	0.836	0.795	0.910	0.823
70	0.330	0.857	0.874	0.815	0.953	0.849
65	0.282	0.890	0.902	0.863	1.017	0.885
60	0.241	0.924	0.935	0.902	1.025	0.920
50	0.174	0.974	1.015	0.978	1.134	0.989
40	0.123	1.042	1.065	1.039	1.227	1.049
30	0.083	1.071	1.113	1.080	1.293	1.088
20	0.050	1.089	1.152	1.132	1.353	1.124
10	0.023	1.106	1.169	1.154	1.366	1.143
5	0.011	1.111	1.177	1.159	1.375	1.149

(*) Estos valores no se incluyen en el promedio final.

cada composición de disolvente y el valor adoptado como definitivo, promedio de los tres indicadores seleccionados, como se justifica posteriormente.

En la Figura 1 se hallan representados frente a la fracción molar de dioxano en el disolvente mixto los valores del parámetro π^* procedentes de los indicadores: 4-nitroanisol, 2-nitroanisol y 4-etilnitrobenceno.

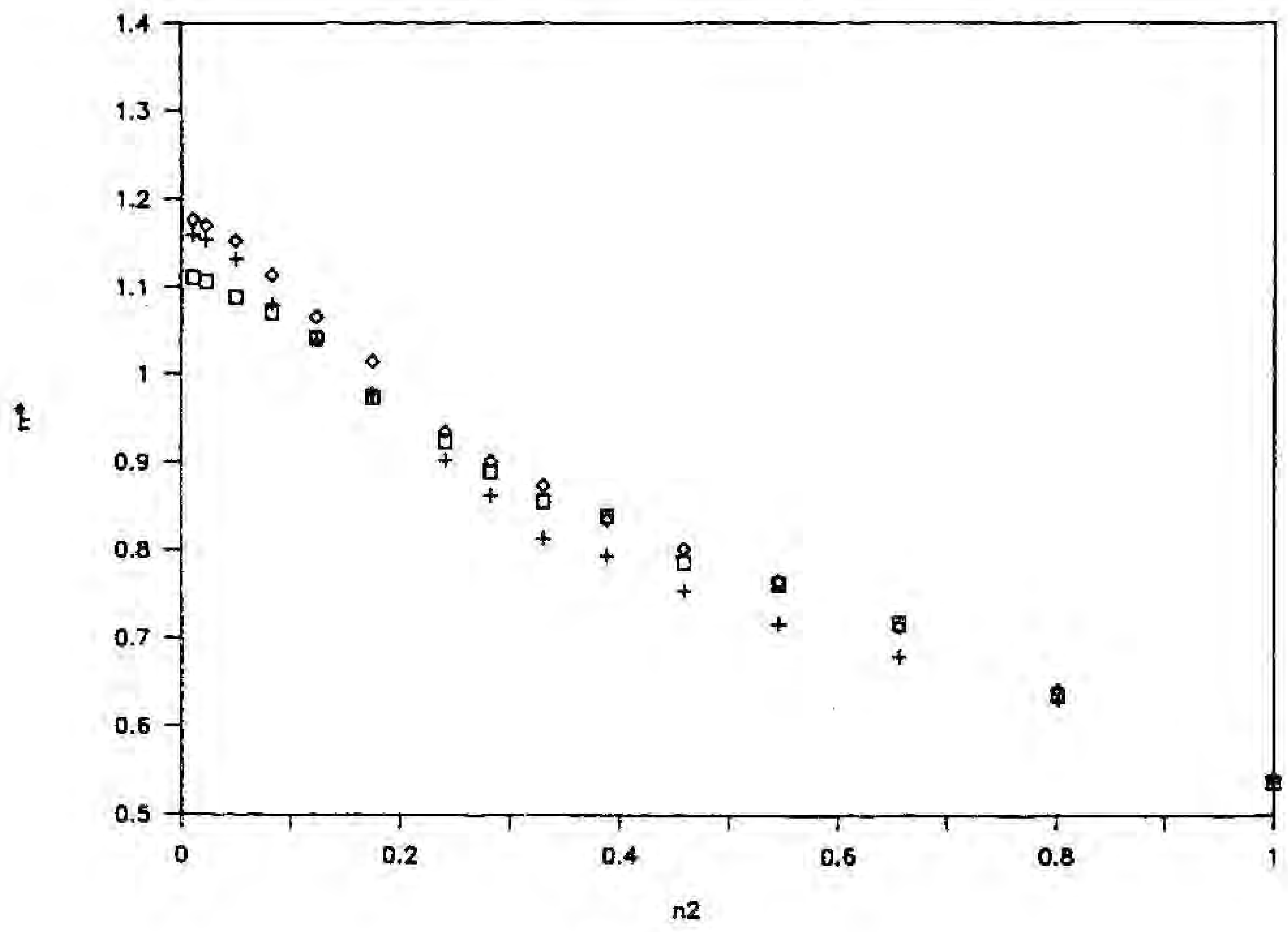


Figura 1

Valores del parámetro π^* obtenidos con los indicadores considerados válidos (◇ 2-nitroanisol, □ 4-nitroanisol y + 4-etilnitrobenceno) frente a la fracción molar de dioxano.

De los cuatro indicadores de referencia escogidos, la N-metil-2-nitroanilina produce resultados que se desvían de forma notoria de los que derivan de los tres restantes. Ostensibles diferencias han sido observadas en otros casos por W. J. Cheong y P. W. Carr (8) quienes han propuesto un criterio para analizar la validez de los diversos indicadores utilizados en la determinación del parámetro π^*, y así justificar la eliminación de resultados aparentemente erróneos. En el caso de los resultados discrepantes de la N-metil-2-nitroanilina se ha aplicado este criterio. Para ello se representan los valores de π^* de cada indicador a cada composición de la mezcla binaria frente a las medias (obtenidas para cada composición de disolvente) de los valores de π^*, procedentes de los diversos indicadores no considerados discrepantes. Cuanto más próximo sea el gráfico obtenido a una recta de pendiente unidad y de ordenada en

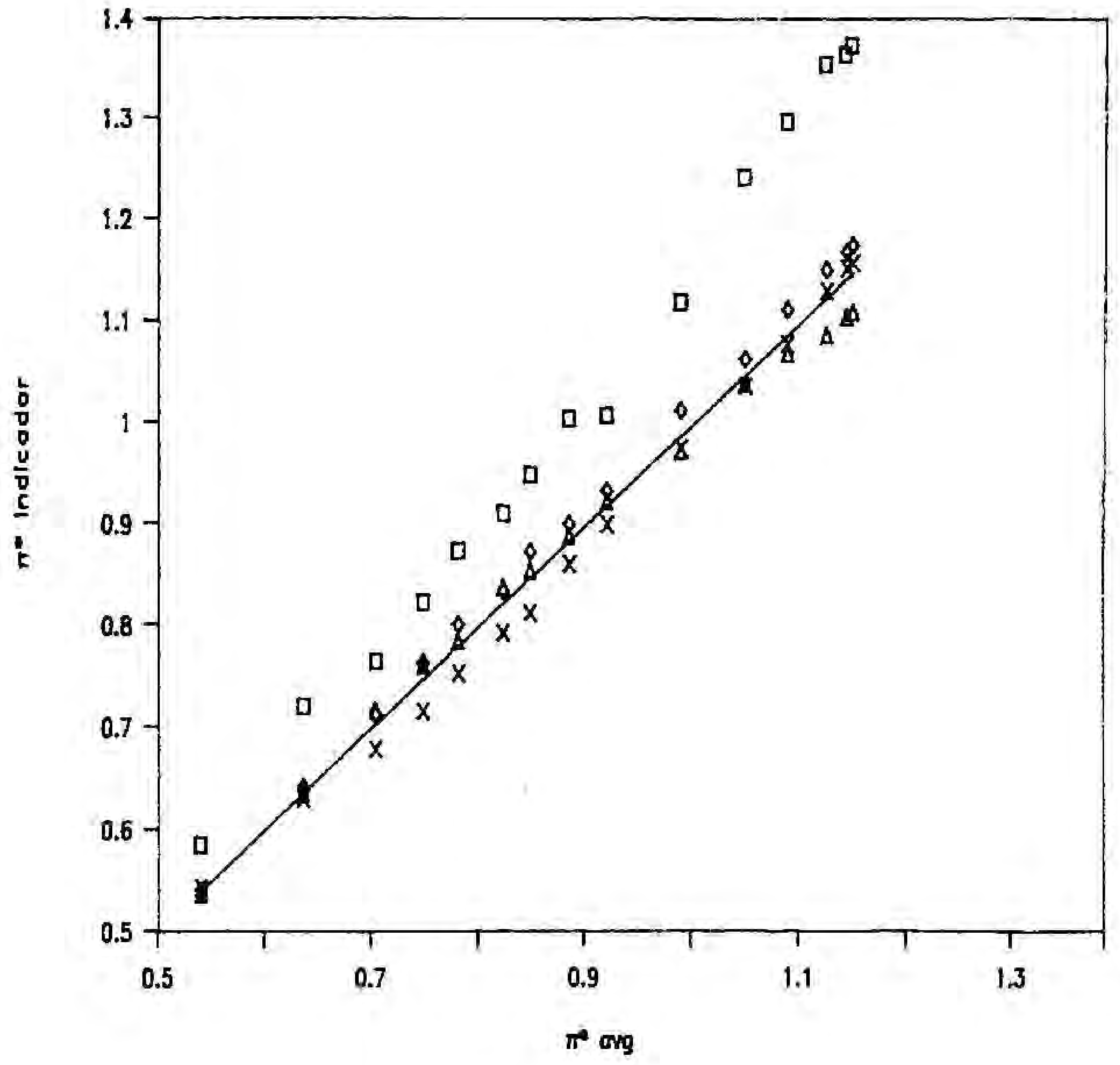


Figura 2

Representación de los valores de π^* procedentes de los cuatro indicadores utilizados (◇ 2-nitroanisol, Δ 4-nitroanisol, × 4-etilnitrobenceno y + N-metil-2-nitroanilina) frente a los valores de π^*, promedio de los indicadores considerados válidos.

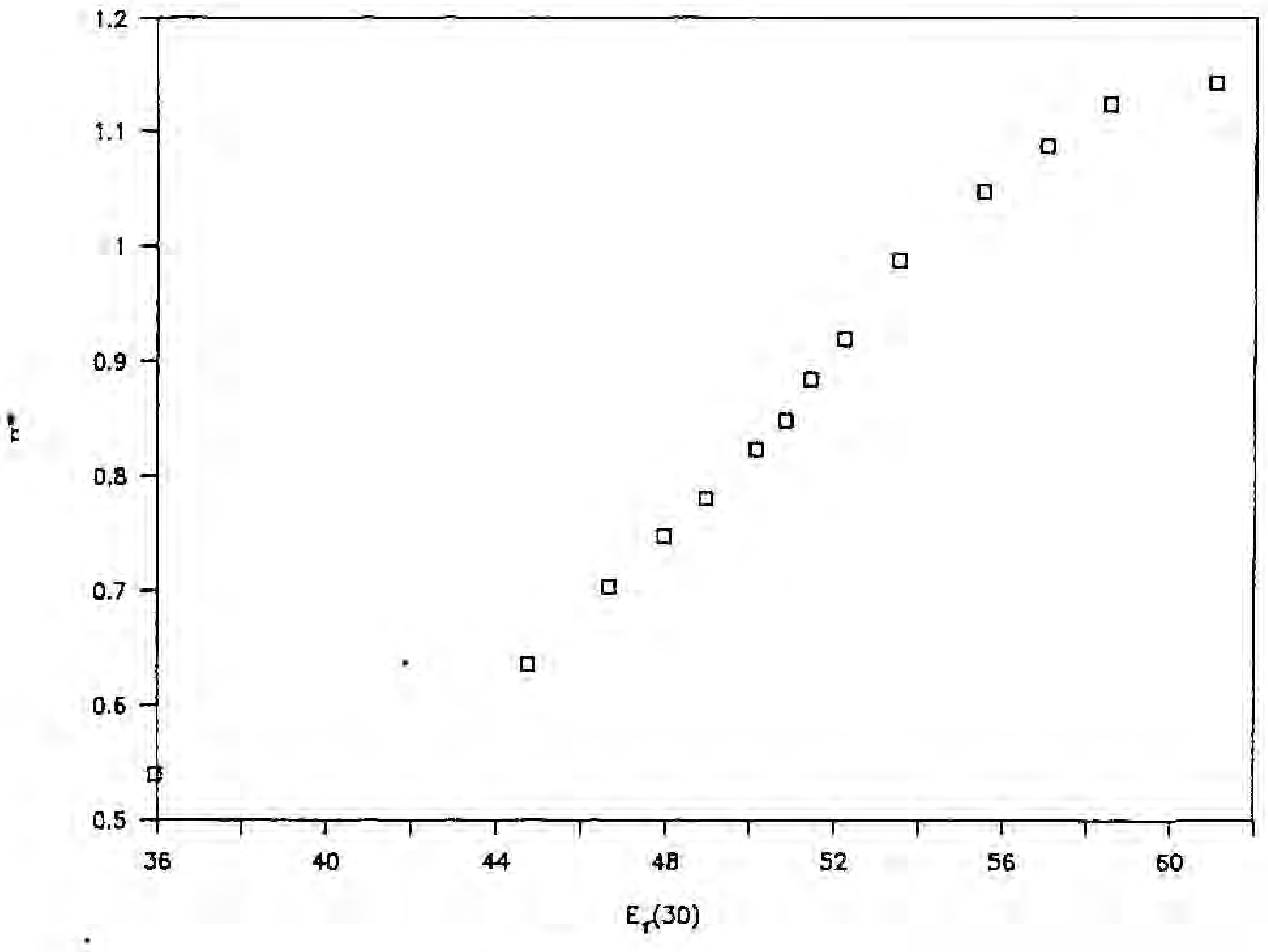


Figura 3

Representación de los valores del parámetro π^* frente a sus homólogos de $E_T(30)$ para las diferentes composiciones de la mezcla dioxano-agua.

el origen igual a cero, tanto más adecuado será el indicador objeto de estudio. La Figura 2 representa los valores de π^* procedentes de los cuatro indicadores respecto al promedio de valores de π^* de los tres indicadores considerados válidos, para cada composición de la mezcla disolvente. La figura manifiesta claramente el carácter discrepante de los resultados que provienen del uso de la N-metil-2-nitroanilina como substancia indicadora. El comportamiento anómalo de este indicador debe ser atribuido a las interacciones de puentes de hidrógeno que establece dicho compuesto con el disolvente.

Si bien todos los solutos de referencia empleados poseen una mínima basicidad que afecta casi imperceptiblemente al valor final del parámetro π^*, (de ahí las pequeñas diferencias entre los resultados que proceden de las diferentes substancias utilizadas y, por tanto, el uso obligatorio de varias de ellas para evitar la influencia en el resultado final de interacciones específicas propias de un determinado compuesto) no es menos cierto que otros parámetros acusan la contribución debida a las interacciones de puentes de hidrógeno en un grado mucho más significativo. Este es el caso del parámetro $E_T(30)$, que, debido a la estructura del compuesto betaínico empleado como indicador solvatocrómico para su medida, sufre un desplazamiento espectral no sólo por el cambio de polaridad del disolvente, sino también por la formación de puentes de hidrógeno con éste, según la capacidad de donación de protones del mismo.

La representación de los valores promedio de π^* frente a sus homólogos correspondientes de $E_T(30)$ para las diferentes mezclas agua-dioxano (Figura 3) demuestra claramente la diferente naturaleza existente entre ambos parámetros, manifestada por una función que dista mucho de la linealidad, reflejo de la falta de analogía entre los desplazamientos espectrales correspondientes a cada una de las variables analizadas.

Fue Kamlet (8) el primero en postular que el parámetro $E_T(30)$ podía expresarse como una combinación lineal de los parámetros α y π^*, definitorios de las contribuciones de interacción por puentes de hidrógeno y de polaridad antes mencionadas. Más tarde W. J. Cheong y P. W. Carr, en sus estudios sobre la naturaleza de diferentes mezclas hidroorgánicas, propusieron la siguiente ecuación general:

$$E_T(30) = 31{,}92\ (\pm 3{,}8) + 11{,}42\ (\pm 5{,}6)\ \pi^* + 15{,}95\ (\pm 3{,}0)\ \alpha \qquad [2]$$

a la que se ajustaban los datos provenientes de todas las mezclas por ellos estudiadas (8).

Para las diferentes mezclas dioxano-agua, tomando los valores del parámetro π^* determinados en el presente trabajo y los valores de la bibliografía para $E_T(30)$ (5, 6) y α (9) se obtienen las siguientes ecuaciones:

— Con los valores de π^* procedentes del 4-nitroanisol:

$E_T(30) = 30{,}79 + 12{,}34\ \pi^* + 17{,}49\ \alpha$ Desv. est. = 0,181 r = 0,9997

— Con los valores de π^* procedentes del 2-nitroanisol:

$E_T(30) = 31{,}80 + 10{,}68\ \pi^* + 17{,}86\ \alpha$ Desv. est. = 0,295 r = 0,9992

— Con los valores de π^* procedentes del 4-etilnitrobenceno:

$E_T(30) = 32{,}66 + 9{,}29\ \pi^* + 19{,}08\ \alpha$ FDesv. est. = 0,399 r = 0,9986

La ecuación final que establece la relación existente entre los parámetros en estudio es la siguiente:

$E_T(30) = 31{,}75 + 10{,}83\ \pi^* + 10{,}06\ \alpha$ Desv. est. = 0,295 r = 0,9982

obtenida a partir de los valores medios de π^* adoptados como definitivos, procedentes de los tres indicadores conjuntamente. Para todas las ecuaciones anteriores dev. est. representa la desviación estándar de los residuales, esto es, de las diferencias entre los valores de $E_T(30)$ estimados a partir de la ecuación de ajuste y sus valores experimentales.

Todas estas expresiones concuerdan con la propuesta por los dos autores ya mencionados, corroborando la diferencia de significado físico de los parámetros que se comparan, π^* y $E_T(30)$.

AGRADECIMIENTO

Los autores quieren expresar su agradecimiento a la C.I.C.Y.T (PB87-0061) y a la C.I.R.I.T. por su ayuda financiera en el desarrollo de este trabajo.

BIBLIOGRAFIA

1. DIMROTH, K. y REICHARDT, C.; *Z. Anal. Chem.*, **215**: 344 (1966).
2. TAFT, R. W. y KAMLET, M. J.; *J. Amer. Chem. Soc.*, **98**, 2886 (1976).
3. KAMLET, M. J. y TAFT, R. W.; *J. Amer. Chem. Soc.*, **98**, 377, (1976).
4. KAMLET, M. J., ABBOUD, J. J. y TAFT, R. W.; *J. Amer. Chem. Soc.*, **99**, 6027 (1977).
5. REICHARDT, C.; «Solvent and Solvent Effects in Organic Chemistry». *VCH* Weinheim, 1988.
6. CASASSAS, E., FONRODONA, G. y DE JUAN, A.; *Inorg. Chim. Acta* (en prensa).
7. VOGEL, A. I.; «A textbook of practical organic chemistry». Longmans Green. 5a. ed. London 1989, pág. 177.
8. CHEONG, W.; J. y CARR, W.; *Anal. Chem.*, **60**, 820 (1988).

Journal of Solution Chemistry, Vol. 21, No. 2, 1992

Solvatochromic Parameters for Binary Mixtures and a Correlation with Equilibrium Constants. Part I. Dioxane-Water Mixtures

E. Casassas,[1,2] G. Fonrodona,[1] and A. de Juan[1]

Received December 5, 1990; Revised July 9, 1991

The values of the solvatochromic parameters α and β were determined at 25°C for dioxane-water mixtures from 0 to 100% of dioxane. These values as well as those of the Reichardt polarity parameter $E_T(30)$ and the polarity-polarizability π are correlated with acid dissociation constants and other equilibrium constants in solvent mixtures of the same composition. As a general rule, two linear zones with different slopes are obtained, one zone covering water-rich solutions, and the other dioxane-rich solutions. The change in behavior takes place at about 55% (v/v) dioxane for all equilibria studied. A fit of pK to an equation of the multiparametric form proposed by Kamlet and Taft shows in most cases a linear dependence on π* alone, in other cases a dependence on π* and β.*

KEY WORDS: Binary mixtures; dioxane-water; Kamlet and Taft α, β and π* parameters; $E_T(30)$ parameter.

1. Introduction

Recently, considerable effort has been devoted to the determination of empirical parameters for the description of the fundamental properties of pure solvents in order to develop expressions for the prediction of the effects of solute-solvent interactions on the properties of solutes. The field of binary solvent mixtures has been much less explored, although research has been reported concerning the variation of some solute properties, for instance, of several equilibrium constants (mostly for the acid-base dissociation of solutes or for complex-forming reactions of some solutes with metal ions) as a function of composition of binary solvent mixtures over limited ranges of solvent composition.

The correlation of these equilibrium constants with several

[1] Department of Analytical Chemistry, University of Barcelona, Diagonal 647, 08028 Barcelona, Spain.

[2] To whom correspondence should be addressed.

composition-depending functions, such as the solvent polarity expressed as a function of bulk properties of the mixture (*e.g.*, the relative permittivity) or of its microscopic parameters, has been attempted in spite of the fact that in solvent mixtures the concept of polarity has a limited meaning. The correlations found have never been fully satisfactory, especially when the composition range covers the full span from 0 to 100%. Since most of the solvent mixtures used in practical applications contain a polar constituent, for which hydrogen-bonding interactions with acid-base solutes or complexing solutes can be important, in the present paper an attempt is made to relate the variation of equilibrium constants with the solvatochromic parameters of Kamlet and Taft which measure the hydrogen-bond interactions α for hydrogen bond-donor and β for hydrogen bond-acceptor solvents, in order to establish the extent to which these parameters will modify known relationships with the polarity parameter π^*.

The use of solvatochromic parameters with mixed solvents has been discussed[1] on the basis that no theoretical meaning can be assigned to these parameters in such environments, since the procedures for their measurement and the microscopic scale at which they are applied involve molecular properties of each constituent of the solvent mixture and the solute studied. The application of the empirical values of solvatochromic parameters to solutes in solvent mixtures implies, then, the assumption that the properties of the solvation shell of certain reference solutes (the solvatochromic indicators) are quite similar to those of any other solute. Actually, it is known that in the solvation shell of any solute in a mixed solvent, a sorting-in of the more polar component (dielectric enrichment) is produced which is mainly determined by the charge or dipole features of the solute molecule. In any case, this apparent oversimplification is much less dangerous than that which tries to define the cybotactic zone of the solute through the bulk properties of the solvent mixture.

Thus, the empirical scales of microscopic parameters can be considered the best tool now available to explain solvent-dependent microscopic processes because they are the ones that reflect most reliably the complete picture of all intermolecular forces acting between solute and solvent molecules. Their usefulness for binary solvent mixtures has been widely confirmed in different fields of chemistry. Reichardt reported many studies on the use of $E_T(30)$ in solvent mixtures,[2] and Langhals worked in the same sense.[3] Recently, this parameter has been applied to binary mixtures to investigate the existence of specific solvation of solutes.[4,5] In a more practical way, $E_T(30)$ has been applied to the interpretation of chromatographic features, such as retention times[6,7] and selectivities.[8]

According to Langhals,[3,9] for binary mixtures of solvents, the relationship between $E_T(30)$ and composition is given by

$$E_T(30) = E_D \ln(c_p/c^* + 1) + E_T(30)_o \qquad (1)$$

where c_p is the molar concentration of the more polar component, E_D and c^* are adjustable parameters which are specific for the binary mixture under study, $E_T(30)$ and $E_T(30)_o$ are the polarity parameters for the mixed solvent and for the pure more-polar component, respectively. Equation (1) is not valid for certain solvent mixtures such as dioxane-water over the full range of miscibility where two regions of linear behavior with different slopes are obtained. The parameter c^* defines the point of transition between them. Furthermore, for dioxane-water mixtures it was observed [10] that (a) the log of the autoprotolysis constant also correlates with $E_T(30)$ through two linear expressions with the same transition point defined by Langhals and (b) the log of the acid dissociation constant for the carboxyl group in 3-hydroxy-2-naphthoic acid (hnca) follows the same behavior and shows the full range of mixture compositions divided into two linear zones with an inflexion point at the critical concentration c^* of about 60% (*v*/*v*) dioxane. In this paper, the general validity of the two linear-segments principle is tested using the protonation constants for several other acids [two carboxylic acids (glycine and hnca) and three ammonium cationic acids (NH_3^+ groups in glycine and ethylenediamine)] and some complex formation constants for several systems [K_1, K_2, and K_3 for the complexes in Ni(II)-glycine and Zn(II)-ethylenediamine systems] in dioxane-water mixtures.

The now widely used multiparametric approach proposed by Kamlet and Taft,[11-13] which separates solute-solvent interactions into pure contributions (hydrogen-bond acidity, hydrogen-bond basicity, polarity-polarizability) has also been applied to binary mixtures.[14] This approach, applied to the microscopic parameters of the mixed solvents used as mobile phases in liquid chromatography, yields correlations useful to rationalize capacity factors[15,16] eluotropic strength[17] and many other chromatographic parameters. The solvatochromic parameters find application also in the same field as measures of acidity, basicity and polarity in order to classify solvent, and have been used by Snyder and coworkers in order to re-evaluate their famous solvent triangle.[18]

The parameters α, β and π^* were determined by Kamlet, Taft and co-workers[19] for many pure solvents but data are scanty for binary mix-

Table I. Spectral Data for the Betaine Dye (2) and the $E_T(30)$ Parameter in Dioxane-Water Mixtures[a]

% Dioxane[b]	n_2[c]	$\nu(2)_{max}$[d]	$E_T(30)$[e]	% Dioxane[b]	n_2[c]	$\nu(2)_{max}$[d]	$E_T(30)$[e]
100	1.000	12.59	36.0	65	0.282	18.00	51.5
95	0.801	15.66	44.8	60	0.241	18.29	52.3
90	0.655	16.33	46.7	50	0.174	18.75	53.6
85	0.545	16.79	48.0	40	0.123	19.45	55.6
80	0.458	17.14	49.0	30	0.083	19.97	57.1
75	0.388	17.57	50.2	20	0.050	20.50	58.6
70	0.330	17.80	50.9	10	0.023	21.37	61.1

[a] All $E_t(30)$ data given are taken from Ref. 5 except those at 95, 75 and 65% dioxane, which are evaluated in the present work by use of the equation: $E_t(30) = h\lambda_{max}\, N_A = h_c N_a / \nu_{max}$. Concentrations of betaine dye used were 1.8×10^{-4}, 3.7×10^{-4} and 4.8×10^{-4} mol-L^{-1}. [b] Volume % dioxane. [c] Mole fraction dioxane. [d] Units: kK. [e] Units: kcal-mol^{-1}.

tures of solvents. In the present work, the values for $E_T(30)$, α_1 and β_2 for some binary mixtures of dioxane-water covering a wide range of compositions were determined. Correlations of these parameters with the values for the protonation constants of some acids and the stability constants of some complexes are explored.

2. Experimental

Dioxane (Probus, r.a. and Merck, p.a.) previously purified by Eigenberger's method was used.[(20)] Water was first deionized and twice distilled over potassium permanganate. The solvatochromic indicators used were: pyridinium N-phenoxide betaine dye (obtained from Prof. E. Bosch), 4-nitroanisole (Merck, previously purified by recrystallization and treatment with carbon black), and 4-nitrophenol (Fluka, r.a). Stock solutions of indicators in dioxane were used to prepare the test solutions at several concentrations of indicator, around $2\times10^{-4}M$, and at the desired solvent composition (by addition of the required amounts of dioxane and water).

In the spectrometric studies a Beckman-DU7 spectrophotometer, equipped with 1 cm quartz cells and attached to an IBM-PC through a RS232 interface, was used. The spectra acquisition was controlled by the Beckman Data Capture Software.

The spectra of the test solutions of each indicator were recorded against a blank consisting of the dioxane-water mixture of identical composition as the test solution. Three replicates were obtained of each

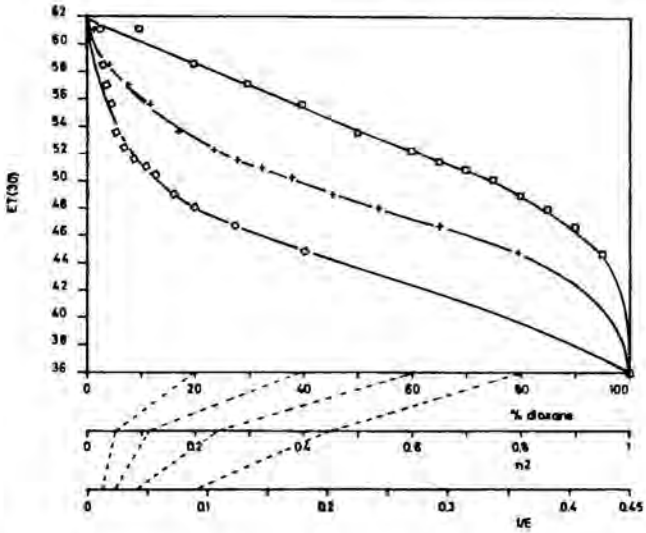


Fig. 1. $E_T(30)$ parameter *vs.* volume fraction of dioxane (□), or *vs.* molar fraction (+) of cosolvent in dioxane-water mixed solvent and *vs.* reciprocal of the dielectric constant (◊) of these solvents. Broken lines unite abscissae values for the same solutions.

Table II. Spectral Data for 4-Nitroanisole(1) and Betaine Dye(2) and the α_1 Parameter in Dioxane-Water Mixtures [a]

% Dioxane	$\nu(1)_{max}$	$\nu(2)_{max}$	α_1	% Dioxane	$\nu(1)_{max}$	$\nu(2)_{max}$	α_1
100	32.86	12.59	-0.07	65	32.03	18.00	0.55
95	32.63	15.66	0.35	60	31.96	18.29	0.57
90	32.43	16.33	0.40	50	31.84	18.75	0.61
85	32.34	16.79	0.45	40	31.68	19.45	0.67
80	32.28	17.14	0.48	30	31.61	19.97	0.74
75	32.15	17.57	0.51	20	31.57	20.50	0.81
70	32.11	17.80	0.54	10	31.53	21.37	0.94

[a] All α_1 values given were evaluated using the equation: $\alpha_1 = \nu_{(2\text{-}1)} / 6.24$ from spectral data reported in the present work. Concentrations of betaine dye were as in Table I, concentrations used of 4-nitroanisole were 3×10^{-5}, 6.6×10^{-5}, and 8.9×10^{-5} mol-L^{-1}. For units see Table I.

spectrum from identical independently prepared test solutions. From the digitalized average spectra (softened according to usual computer programs) the wavelength of the absorption maximum was evaluated.

For every solvent composition this procedure was followed at three different concentration levels of the indicator. The gross average

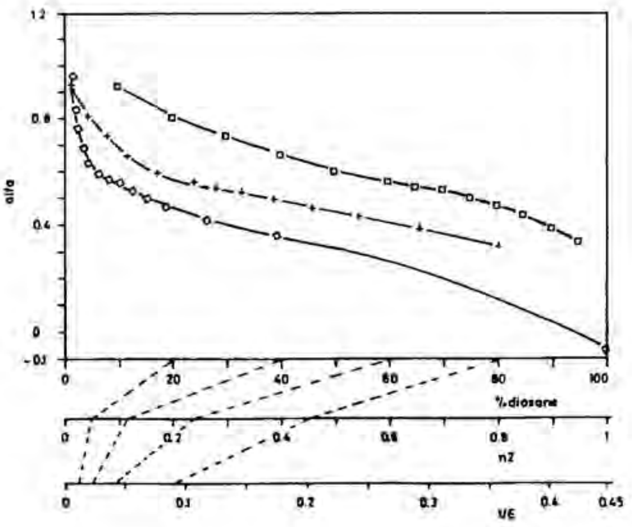


Fig. 2. α_1 parameter *vs.* volume fraction of dioxane (□), or *vs.* molar fraction (+) of cosolvent in dioxane-water mixed solvent and *vs.* reciprocal of the dielectric constant (◊) of these solvents. Broken lines unite abscissae values for the same solutions.

Table III. Spectral Data for 4-Nitroanisole (4) and 4-Nitrophenol (3) and the β_2 Parameter in Dioxane-Water Mixtures[a]

% Dioxane	$\nu(4)_{max}$	$\nu(3)_{max}$	β_2	% Dioxane	$\nu(4)_{max}$	$\nu(3)_{max}$	β_2
100	32.86	32.80	0.419	60	31.96	31.58	0.596
95	32.63	32.10	0.632	50	31.84	31.52	0.575
90	32.43	31.92	0.636	40	31.68	31.49	0.525
85	32.34	31.82	0.640	30	31.61	31.50	0.494
80	32.28	31.77	0.638	20	31.57	31.51	0.474
75	32.15	31.71	0.613	10	31.53	31.57	0.432
70	32.11	31.64	0.628	5	31.52	31.57	0.428
65	32.03	31.61	0.610				

[a] All β_2 values given were evaluated using the equations: $\beta_2 = -\nu_{(3\text{-}4)} / (2.80)(0.825)$, from spectral data obtained in the present work. Concentration of 4-nitroanisole were as given in Table II; concentrations of 4-nitrophenol were: 2×10^{-6}, 2×10^{-5}, and 4×10^{-5} mol-L^{-1}. For units see Table I.

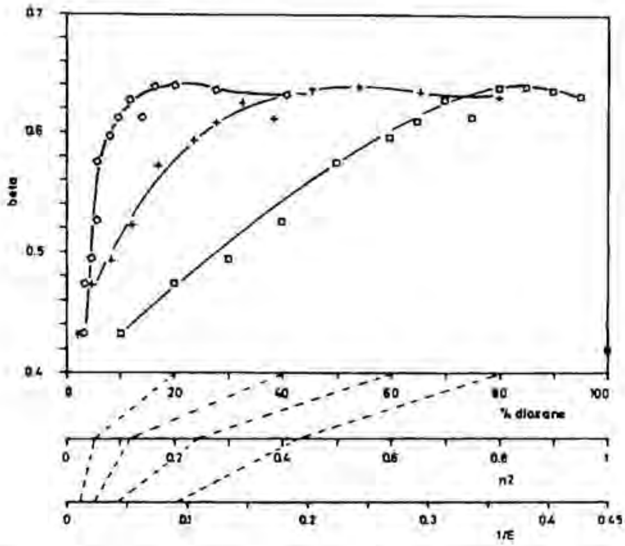


Fig. 3. β_2 parameter *vs.* volume fraction of dioxane (□), or *vs.* molar fraction (+) of cosolvent in dioxane-water mixed solvent and *vs.* reciprocal of the dielectric constant (◊) of these solvents. Broken lines unite abscissae values for the same solutions.

of the wavelengths of maximum absorption at each solvent composition was taken as the final value for the calculation of the solvatochromic parameters.

3. Results

In Table I the $E_T(30)$-values for several dioxane-water mixtures (from the literature[2] or evaluated in the present work) are shown. The dependence of $E_T(30)$ on the volume fraction and mole fraction of this dioxane and on the reciprocal of the dielectric constant of the mixed solvent are shown in Fig. 1.

In Tables II and III the α_1 and β_2 values, respectively, are given for several dioxane-water mixtures. The dependences of these parameters on the same composition-dependent variables as for $E_T(30)$ are depicted in Figs. 2 and 3.

For dioxane-water mixed solvents, the ionic product K_w of the medium were determined potentiometrically by the Gran[10] procedure. The dependence of pK_w on α_1 and β_2 is shown in Fig. 4.

Correlation of the protonation constants of the carboxyl group in glycine[21] and in hnca acid[10] and of the NH_3^+ group in glycine and ethylenediamine.[21] in dioxane-water mixtures with the $E_T(30)$, α_1 and

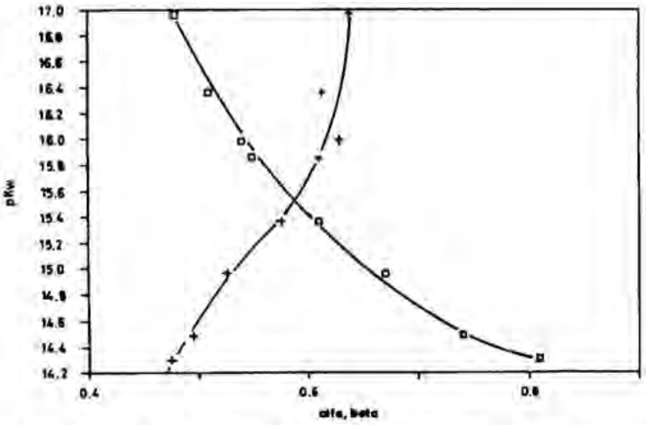


Fig. 4. The reciprocal logarithm of the ionic product in dioxane-water mixtures as a function of the α (□) and β (+) parameters.

β_2 values found in this work are shown in Fig. 5. Stability constants for Ni(II)-glycine and Zn(II)-ethylenediamine complexes in dioxane-water solutions at several solvent compositions[21] also were correlated with $E_T(30)$, α_1 and β_2 values found in the present work and are shown in Fig. 6.

4. Discussion

It has been observed for all acid-base equilibria considered that the correlation of the acid-dissociation constants with $E_T(30)$ yields two linear segments (Fig. 5a), similar to what was observed earlier with hnca.[10] The linearity of the two segments, as well as their intercept, are well defined in the case of carboxylic acids but have a poorer definition for the deprotonation of ammonium groups. The intercept of both linear segments occurs at about 56% dioxane (*v*/*v*) (n_2 = 0.212), the exact values are given in Table IVb. The dependence of all these acid-base equilibrium constants as well as those of the ionic product of the medium on α_1 or on β_2 follow analogous patterns (Fig. 5b, 5c and 4), the β_2 correlation being perhaps less clear as the regression coefficient values in Table IVa show.

The trends observed for the formation constants of binary complexes in dioxane-water solutions show a similar pattern to that described for acid dissociation constants: two linear segments appear in their plots *vs.* $E_T(30)$, α_1 or β_2 (Figs. 6a-c), with an inflection point at about the same solvent composition as before, as given in Table V, where the parameters for the linear equations and their regression coef-

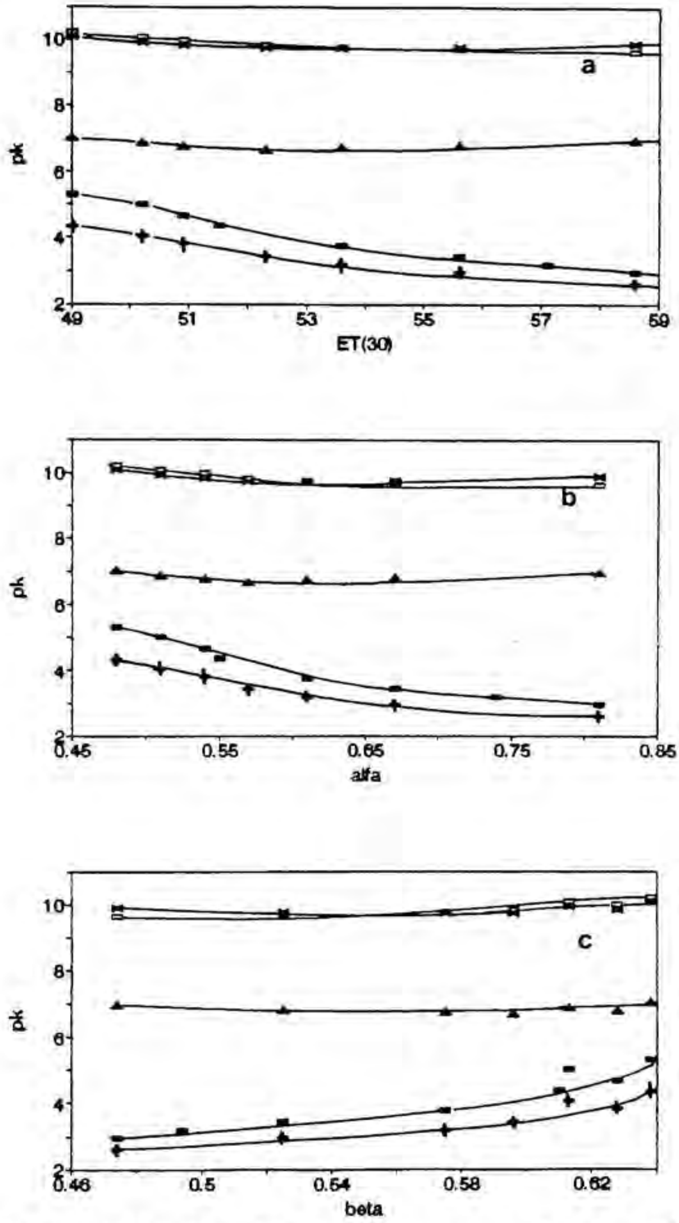


Fig. 5. The reciprocal logarithm of the dissociation constants pK_a for hnca (■), gly (+,□) and en (▲,x) for the different dioxane-water mixtures are plotted *vs.* the $E_T(30)$ (a), α_1 (b), and β_2 (c) parameters.

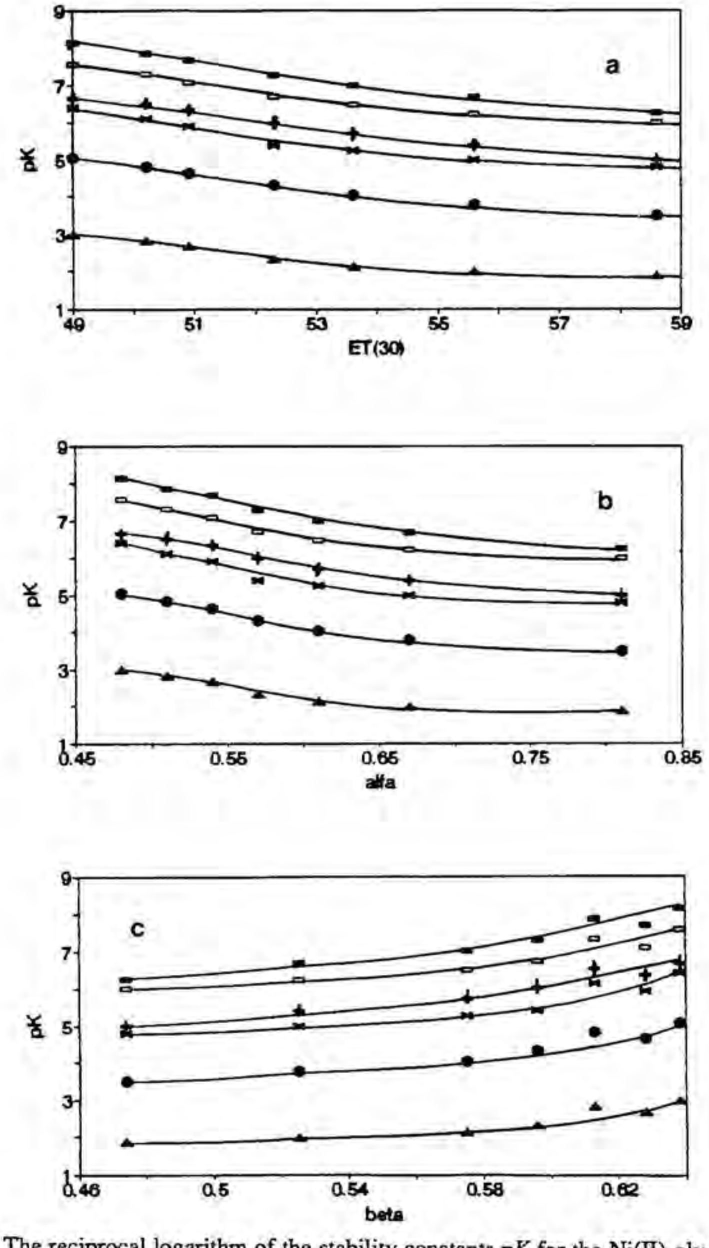


Fig. 6. The reciprocal logarithm of the stability constants p*K* for the Ni(II)-gly (■, +, •) and Zn(II)-en (□, x, ▲) for the different dioxane-water mixtures are plotted *vs.* the $E_T(30)$ (a), α_1 (b), and β_2 (c) parameters.

Table IVa. Parameters in the Linear Equations for the Acid Dissociation Constants of Selected Compounds[a]

	Water-Rich Zone[b]			Dioxane-Rich Zone[c]		
System	*s*	*I*	*C.C.*	*s*	*I*	*C.C.*
			$E_T(30)$			
hnca	0.162(5)	12.44(2)	0.999	0.36(4)	23.15(8)	0.985
gly_1	0.114(1)	9.276(1)	0.999	0.27(1)	17.58(3)	0.998
gly_2	0.023(3)	11.02(1)	0.993	0.12(1)	16.01(1)	0.997
en_1	0.05(1)	3.98(3)	0.984	0.1(2)	11.84(4)	0.974
en_2	0.03(1)	7.97(2)	0.986	0.10(1)	15.28(3)	0.986
			α_1			
hnca	3.9(3)	6.14(5)	0.992	12.0(1)	11.24(8)	0.985
gly_1	2.7(3)	4.83(5)	0.993	9.6(5)	8.95(4)	0.996
gly_2	0.5(1)	10.09(2)	0.974	4.2(3)	12.23(2)	0.995
en_1	1.27(8)	5.95(1)	0.998	3.6(5)	8.71(3)	0.979
en_2	0.81(2)	9.23(1)	0.998	3.8(4)	11.90(2)	0.991
			β_2			
hnca	0.70(5)	7.8(6)	0.995			
gly_1	5.69(2)	0.078(1)	0.999			
gly_2	1.1(1)	9.08(1)	0.993			
en_1	2.5(4)	8.16(3)	0.984			
en_2	1.6(2)	10.65(1)	0.986			

[a] $\log K = sX + I$ [where $X = E_T(30)$, α_1 or β_2] (hnca, 3-hydroxy-2-naphthoic acid; gly, glycine; en, ethylenediamine); values in parentheses are standard deviation in the last significant figure; *C.C.*, correlation coefficient. [b] Four data are used to do the regression analysis for every substance. [c] Four data are used to do the regression analysis for hnca; three data are used for the rest of substances.

ficients are also given.

These results confirm the existence of a change in the nature of the solute-solvent interactions in dioxane-water mixtures which occurs around 55% volume fraction or 0.205 mole fraction as was previously described in a study on the acid-base behavior of hnca.[10] The change in behavior as postulated by Langhals[3] has to be a consequence of a change in the solvent structure.

The mixed solvent under study is amphiprotic in nature as far as the formation of hydrogen bonds is concerned: the water component has strong hydrogen bond donor (HBD) ability, and both components show important hydrogen bond acceptor (HBA) properties. Then, the acid-

Table IVb. Abscissae of the Intercept of the Two Linear Segments[a]

System	$X = E_T(30)$	$X = \alpha$	$X = \beta^b$
hmca	53.3	0.609	0.565
gly_1	53.1	0.602	0.576
gly_2	52.7	0.589	0.584
en_1	52.5	0.572	0.589
en_2	52.6	0.579	0.586
Average Value[c]	52.8	0.590	0.580

[a] See Table IVa for abbreviations. [b] The linear equation for the dioxane-rich zone is not given in Table IVa because of a rather bad correlation coefficient. [c] These average values belong to solutions with the following compositions: 56% dioxane (n_2 = 0.212) for $X = E_T(30)$; 55% dioxane (n_2 = 0.205) for $X = \alpha$ or β.

base or complex-forming properties of solutes should be related to the microscopic parameters of the solvent mixture through an equation of the general linear-combined type used by Kamlet and Taft

$$X_s = X_o + a\alpha + b\beta + s\pi^* + \ldots \qquad (2)$$

where X is the concerned property and a, b, and s are constants which measure the susceptibility of this property to changes in HBD acidity, HBA basicity and solvent polarity, respectively.

Thus, a correlation of the whole set of data with the microscopic parameters according to the general expression has been stepwise tested. In the first place, only the two parameters α_1 and β_2 have been used and, since the resulting correlation was poor, values of π^*[(22)] had to be introduced in the linear regression analysis. Then, a rigorous statistical treatment was applied to find out which multiparametric approach is the best to explain the variation of each equilibria in the full range of mixture compositions. It is well known that the fit quality of a linear regression cannot be ensured only by a high value of its correlation coefficient (*C.C.*),[(23)] since this parameter can be increased by the inclusion of unnecessary variables which overfit the set of data.

The procedure used in this work to determine which set of independent variables provides the best description of the behavior of each one of the data sets studied begins with an attempt of fit using the most extended form of the Kamlet-Taft equation, which includes all the possible variables that could explain the variation of the complex formation constants or protonation constants. The program used yields the cor-

Table Va. Parameters in the Linear Equations for the Complexation Constants in Selected Systems[a]

	Water-Rich Zone[b]			Dioxane-Rich Zone[c]		
System	s	I	$C.C.$	s	I	$C.C.$
			$E_T(30)$			
A, K_1	0.15(1)	14.86(3)	0.998	0.26(1)	20.88(1)	0.999
A, K_2	0.14(1)	13.13(3)	0.998	0.20(2)	16.50(6)	0.985
A, K_3	0.11(1)	10.00(2)	0.998	0.22(1)	15.65(2)	0.999
B, K_1	0.09(1)	11.51(5)	0.990	0.26(1)	20.45(2)	0.998
B, K_2	0.09(1)	10.17(5)	0.988	0.30(2)	21.13(6)	0.993
B, K_3	0.05(1)	4.94(3)	0.984	0.20(2)	13.14(5)	0.991
			α			
A, K_1	3.5(6)	9.11(8)	0.986	9.2(9)	12.53(6)	0.989
A, K_2	3.4(6)	7.70(8)	0.984	7.0(1)	10.07(7)	0.976
A, K_3	2.7(5)	5.65(6)	0.986	7.6(6)	8.72(4)	0.993
B, K_1	2.3(6)	6.55(8)	0.968	9.3(8)	12.04(5)	0.993
B, K_2	2.2(6)	6.55(8)	0.965	10(1)	11.5(1)	0.979
B, K_3	1.3(3)	2.87(5)	0.958	7(1)	6.5(7)	0.976
			β			
A, K_1	7.3(4)	2.78(3)	0.998	16(7)	2.7(2)	0.859
A, K_2	6.9(3)	1.72(2)	0.999	12(6)	1.4(2)	0.826
A, K_3	5.5(3)	0.87(2)	0.998	13(6)	3.7(2)	0.838
B, K_1	4.7(1)	3.73(1)	0.999	16(7)	3.1(2)	0.836
B, K_2	4.6(2)	2.57(1)	0.999	19(8)	6.2(2)	0.863
B, K_3	2.7(2)	0.58(1)	0.997	13(5)	5.6(2)	0.862

[a] $\log K = sX + I$ [where $X = E_T(30)$, α_1 or β_2]. A, Ni(II)-gly system; B, Zn(II)-en system); values in parentheses are standard deviation in the last significant figure; $C.C.$ = correlation coefficient. [b] Four data are used to do the regression analysis for every substance. [c] Three data are used to do the regression analysis for every substances.

relation coefficient and the values of a, b, and s, with their associated errors, and furthermore, it checks through application of the t-test, whether these constants are significantly different from zero,[(24)] and only such constants were retained. The constants that are given in Table IV yield the best fits using these criteria.

From the equations obtained for all the complex-forming reactions (independently of the nature of the metal and the ligand) an identical behavior is observed. The variation of the stability constants in the whole range of compositions of the water-dioxane mixtures can be ex-

Table Vb. Abscissae of the Intercept of the Two Linear Segments[a]

System	$X = E_T(30)$	$X = \alpha$	$X = \beta$
Ni(II)-gly_1	53.28	0.610	0.576
Ni(II)-gly_2	54.89	0.640	0.557
M(II)-gly_3	53.65	0.616	0.570
Zn(II)-en_1	53.13	0.602	0.577
Zn(II)-en_2	52.75	0.592	0.585
Zn(II)-en_3	53.40	0.608	0.576
Average Value[b]	53.51	0.611	0.573

[a] See Table Va for abbreviations. [b] These average values belong to solutions with the following compositions: 55% dioxane ($n_2 = 0.205$) for $X = E_T(30)$; 53% dioxane ($n_2 = 0.192$) for $X = \alpha$, and 54% dioxane ($n_2 = 0.198$) for $X = \beta$.

plained by means of the reduced Kamlet and Taft equation

$$pK = pK_o + s\pi^* \tag{3}$$

where the microscopic polarity π^* appears as the solvent property responsible for the changes in the complexation equilibria.

Analogous conclusions can be deduced from the linear relationships obtained for the protonation constants in the cases of acidic solutes with carboxylic groups and of the amino group of glycine. A different expression is reached for the two constants of ethylenediamine, in which the hydrogen-bond basicity of the solvent (β) must be included in the multiparametric equation, which now becomes

$$pK = pK_o + b\beta + s\pi^* \tag{4}$$

Even though the definitive fit is statistically doubtful because the pK data show only a slight variation along the full range of compositions, the result obtained is significant. The introduction of a new variable β can be understood because of the positive charge of the protonated species which could enhance the interactions with the hydrogen-bond acceptor groups of the solvent mixture. Since the pK data used in this work are the only data available that cover an extensive range of solvent composition, the small number of data in each of the structural zones of the solvent do not justify a separate application of multiparametric fit.

Table VI. Values of the Coefficients of the Kamlet and Taft General Expression[a]

	Full Solvent Composition Range[b]				
System	a	b	s	C	$C.C.$
		Acid-Base			
hnca			-6.6(2)	10.36	0.996
gly-K_1			-4.9(2)	8.04	0.993
gly-K_2			-1.6(2)	11.41	0.96
en-K_1		-6(2)	-3(1)	13.65	0.820
en-K_2		-6(2)	-3.3(9)	16.43	0.89
		Complex Formation			
Ni-gly-K_1			-5.3(1)	12.20	0.9980
Ni-gly-K_2			-4.80(8)	10.42	0.9992
Ni-gly-K_3			-4.4(1)	8.42	0.9990
Zn-en-K_1			-4.5(3)	10.98	0.99
Zn-en-K_2			-4.6(1)	9.89	0.98
Zn-en-K_3			-3.4(3)	5.58	0.98

[a] General expression: $pK_{(solv)} = C + a\alpha + b\beta + s\pi^*$ obtained for the acid-base and complex systems studied, in the full solvent composition range. [b] Eight data are used to do this regression analysis for hnca; seven data are used for the rest of substances.

5. Conclusion

Equilibrium constants for some acid-base and complex-forming reactions in water-dioxane mixtures show a different linear relationship to $E_T(30)$ (taken as a measure of solvent polarity) for solvents with dioxane contents richer or poorer than 55-50% (*v*/*v*) (mole fraction = 0.2, implying a molar dioxane-water ratio of about 1:4).

When the equilibrium constants *vs.* some of the solvatochromic parameters over the full range of compositions are subjected to Kamlet and Taft's general equation by an iterative fitting process, in all cases linear functions of only the polarity-polarizability parameter π^* are obtained, except for the deprotonation of ethylenediamine which requires a multiple linear regression involving the two parameters π^* and β. In all cases, the variation of the equilibrium constants is described well by the microscopic properties of the solvent in the cybotactic zone of the solutes.

Acknowledgment

This research was supported by the Spanish Comision Interministerial de Ciencia y Tecnologia (C.I.C.Y.T) Grant Number PB87-0061. We also thank Prof. E. Bosch for her cooperation.

References

1. P. Suppan, *J. Photochem. Photobiol.* **50**, 293 (1990).
2. C. Reichardt, in *Molecular Interactions*, Vol. 3, H. Ratajczak and W. J. Orville-Thomas, eds., (Wiley, Chichester, 1982) p.241.
3. H. Langhals, *Angew. Chem. Int. Ed. Engl.* **21**, 724 (1982).
4. J. R. Haak and J. B. F. N. Engberts, *Rec. Trav. Chim. Pays-Bas.* **105**, 307 (1986).
5. J. G. Dawber, J. Ward, and R. A. Williams, *J. Chem. Soc. Faraday Trans. I* **84**, 713 (1988).
6. B. P. Johnson, M. G. Khaledi, and J. G. Dorsey, *Anal. Chem.* **58**, 2354 (1986).
7. J. J. Michels and J. G. Dorsey, *J. Chrom.* **457**, 85 (1988).
8. B. P. Johnson, M. G. Khaledi, and J. G. Dorsey, *J. Chrom.* **384**, 221 (1987).
9. H. Langhals, *Nouv. J. Chim.* **5**, 97, 511, (1981); **6**, 265 (1982).
10. E. Casassas and G. Fonrodona, *J. Chim. Phys.* **86**, 391 (1989).
11. R. W. Taft and M. J. Kamlet, *J. Am. Chem. Soc.* **98**,2886 (1976).
12. M. J. Kamlet and R. W. Taft, *J. Am. Chem. Soc.* **98**, 377 (1976).
13. M. J. Kamlet, J. L. Abboud, and R. W. Taft, *J. Am. Chem. Soc.* **99**, 6027 (1977).
14. W. J. Cheong and P. W. Carr, *Anal. Chem.* **60**, 820 (1988).
15. P. C. Sadek, P. W. Carr, R. M. Doherty, M. J. Kamlet, R. W. Taft, and M. H. Abraham, *Anal. Chem.* **57**, 2971 (1985).
16. M. J. Kamlet, M. H. Abraham, P. W. Carr, R. M. Doherty, and R. W. Taft, *J. Chem. Soc. Perkin Trans. II* 2087 (1988).
17. J. H. Park and P. W. Carr, *J. Chrom.* **465**, 123 (1989).
18. S. C. Rutan, P. W. Carr, W. J. Cheong, J. H. Park, and L. R. Snyder, *J. Chrom.* **463**, 21 (1989).
19. M. J. Kamlet, J. L. Abboud, and R. W. Taft, *Prog. Phys. Org. Chem.* **13**, 485 (1981).
20. A. I. Vogel, *A Text Book of Practical Organic Chemistry* (Longmans Green, London, 1989), p. 407.
21. K. K. Mui and W. A. E. McBryde, *Can. J. Chem.* **52**, 1821 (1974).
22. E. Casassas, G. Fonrodona, and A. de Juan, *Anales de Química* **87**, (1991) (in press).
23. *Analytical Methods Committee Analyst* **113**, 1469 (1988).
24. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection* (Wiley, New York, 1981).

5.2. Modelling of solvent-dependent processes in water dioxane mixtures, proposals for the establishment of Linear Free Energy Relationships (LFER).

Chemometrics and Intelligent Laboratory Systems, 12 (1991) 29–38
Elsevier Science Publishers B.V., Amsterdam

Assessment of solvent parameters and their correlation with protonation constants in dioxane–water mixtures using factor analysis

E. Casassas *, G. Fonrodona, A. de Juan and R. Tauler

Department of Analytical Chemistry, University of Barcelona, Diagonal 647, 08028 Barcelona (Spain)

(Received 3 September 1990; accepted 25 April 1991)

Abstract

Casassas, E., Fonrodona, G., de Juan, A. and Tauler, R., 1991. Assessment of solvent parameters and their correlation with protonation constants in dioxane–water mixtures using factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 12: 29–38.

The effect of the composition of dioxane–water mixtures on wavelength shifts of different solvatochromic indicators and on pK_a values of different carboxylic acids are studied using factor analysis. Target factor analysis is subsequently applied to find out which are the physicochemical parameters affecting the observed data variation. For wavelength shifts of the selected solvatochromic indicators three main sources of variation are found: solvent polarity, π^*, the hydrogen-bond donor acidity of solvent, α, and the hydrogen-bond acceptor basicity of the solvent, β. For pK_a values of carboxylic acids the main source of variation is identified as being the polarity of the mixed solvent.

INTRODUCTION

Factor analysis (FA) has been used in the study of pure solvent effects on certain physicochemical properties [1–3], but it has not been applied to a similar study for binary mixtures of solvents. A generally accepted assumption for these systems is that every property, e.g. pK_a for acidic substances, or spectroscopic data for absorbing compounds, depends on the concentration ratio of constituents in the mixed solvent.

Solvent effects on some properties of solutes can be quantified [4] by the use of macroscopic parameters such as n_2 (molar fraction) or ϵ (dielectric constant) or microscopic parameters, such as Y (solvent ionizing power) [5], or Z (polarity parameter) [6], or the more commonly used $E_T(30)$, proposed by Dimroth and Reichardt [7] as a measure of polarity, or α, β and π^*, proposed by Kamlet and Taft [8] to quantify the hydrogen-bond donor acidity (α), hydrogen-bond acceptor basicity (β) and the polarity–polarizability of the solvent (π^*). In the present work the real meaning of certain solvent microscopic parameters as applied to water–dioxane mixtures is assessed. In spite of some doubts that have been cast upon the physical significance of solvatochromic parameters in general when they are applied to solvent mixtures [9], these parameters are useful operational tools which have been applied to correlate several solute properties with mixed solvent composition [10].

Microscopic parameters for a given solvent are calculated from the frequencies of the main spectral bands of certain reference solutes, the exact position of these bands depending on the solute–solvent interaction in the cybotactic zone.

Reference substances have been proposed whose spectral shifts are supposed to be sensitive only to the property which has to be measured (e.g., polarity) or to a linear combination of some of these properties (e.g., polarity, hydrogen-bond effects). Whether this is the case or some strange factor affects the value of the shift can be shown by using FA since this method enables the prediction of how many causes of variation include the original data without assuming previously any model nor any kind of behaviour. Once the number of factors of which the spectral data are dependent is known, the nature of these factors can be determined through target factor analysis (TFA). Among the targets taken in consideration in the present work are the microscopic parameters of the solvent.

The same chemometric techniques, FA and TFA, and the same targets are used in a second part of the work to interpret the variation of pK_a of several selected solutes in water–dioxane mixtures.

The fact that these parameters have never been assessed before by using these 'hypothesis-free' methods (FA), the recognized importance of these microscopic parameters to get over the lack of information about the properties of the solvation sphere, very different from those of the bulk solvent, especially for solvent mixtures, and the improvement that the use of these parameters introduces in the explanation of complex microscopic solvent-dependent processes, such as proton transfer equilibria in solvent mixtures, justifies the objectives of the present work.

The present work introduces an 'abstract' way to deal with the spectral shifts of the substances adopted as solvatochromic indicators when solved in a mixed hydroorganic solvent, in order to evaluate the degree of purity of the measures that they afford of every given mixed solvent property. The same hypothesis-free method is applied to the study of the variation of acid–base properties of some selected solutes. To reach the goals above, two sets of data are first analyzed using FA [1] in order to determine the number of independent parameters affecting the data. One set consists of the wavelength shifts of different solvatochromic indicators and the other one of the pK_a values of carboxylic acids in the same solvent mixtures. Target factor analysis, TFA [11], is afterwards applied to the same data to identify the independent parameters, to check the validity of some previously accepted physical models and, eventually, to propose new ones.

EXPERIMENTAL

Reagents

Dioxane (PROBUS, r.a.) purified by Eigenberger's method [12]. Water deionized and distilled twice over potassium permanganate. 4-Nitroanisole (Merck, z.s.) purified by activated carbon treatment in an acetone solution and recrystallization in water. 2-Nitroanisole (Aldrich, r.a.). 4-Ethylnitrobenzene (Aldrich, r.a.). N-Methyl-2-nitroaniline (Aldrich, r.a.). 4-Nitrophenol (Fluka Chemie puriss.). 4-Nitroaniline (Carlo Erba, r.a.). 2,6-Diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenolate (obtained from Prof. E. Bosch, whose kind co-operation is acknowledged). Potassium hydroxide (Merck, r.a.), CO_2-free solution in dioxane–water, prepared by the ion-exchange procedure [13]. Nitric acid (Merck, r.a.). Potassium nitrate (Merck, r.a.). Salicylic acid (Merck, r.a.) purified by sublimation. Propionic acid (Merck, r.a.). 3-Hydroxynaphtalene-2-carboxylic acid (Merck, r.s.) purified by activated carbon treatment and recrystallization from ethyl alcohol. Potassium chloride (Merck, r.a.).

Apparatus

Beckman DU-7 spectrophotometer interfaced (RS232) to an IBM personal computer. Spectral acquisition was controlled through Beckman data capture software. ORION SA 720 potentiometer (precision ±0.1 mV). ORION 90-05 AgCl/Ag reference electrode with ceramic junction and internal reference solution of saturated KCl in the working hydroorganic mixture. ORION 91-01 glass electrode. Double-jacketed cell thermostated at (25 ± 0.1) °C. Metrohm 665 Dosimat autoburette (precision 0.01 ml) with an exchange unit of 5 cm^3 with an antidiffusion burette tip. Magnetic stirrer. The complete titration set-up is connected to a PC computer (HP Vectra ES/12 or HP 9133) through an HP 3421A interface, which allows the full automation of the titration process.

Experimental procedure

Assessment of solvent parameters

For the determination of microscopic parameters of solvents, Kamlet and Taft [8] proposed the use as indicators of several substances whose absorption spectra are sensitive only to the solvent properties to be measured and which are free from contributions produced by other causes. For amphiprotic solvents or amphiprotic mixtures, they proposed the following substances: 2-nitroanisole, 4-nitroanisole, 4-ethylnitrobenzene and *N*-methyl-2-nitroaniline (all of them showing spectral shifts sensitive only to the polarity of the solvent), 2,6-diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenolate (betaine) (the spectral shifts of which are sensitive to the polarity and to the hydrogen-bond donor acidity of the solvent), 4-nitrophenol and 4-nitroaniline (with spectral shift sensitive to the polarity and to the hydrogen-bond acceptor basicity of solvent). All of them are used in the present work, but for the case of betaine several of its frequency values are taken from the literature [14].

The spectra of the test solutions of each indicator are recorded against a blank consisting of the pure solvent (a dioxane–water mixture of identical composition to the test solution). Three replicates are obtained of each spectrum from identical, independently prepared test solutions. The wavelength of the longest-wavelength absorption maximum is determined from the digitized average spectra. For every solvent composition this procedure is followed at three different concentration levels of the indicator. The gross average of the wavelengths of maximum absorption at each solvent composition is taken as the final value for the calculation of solvatochromic parameters.

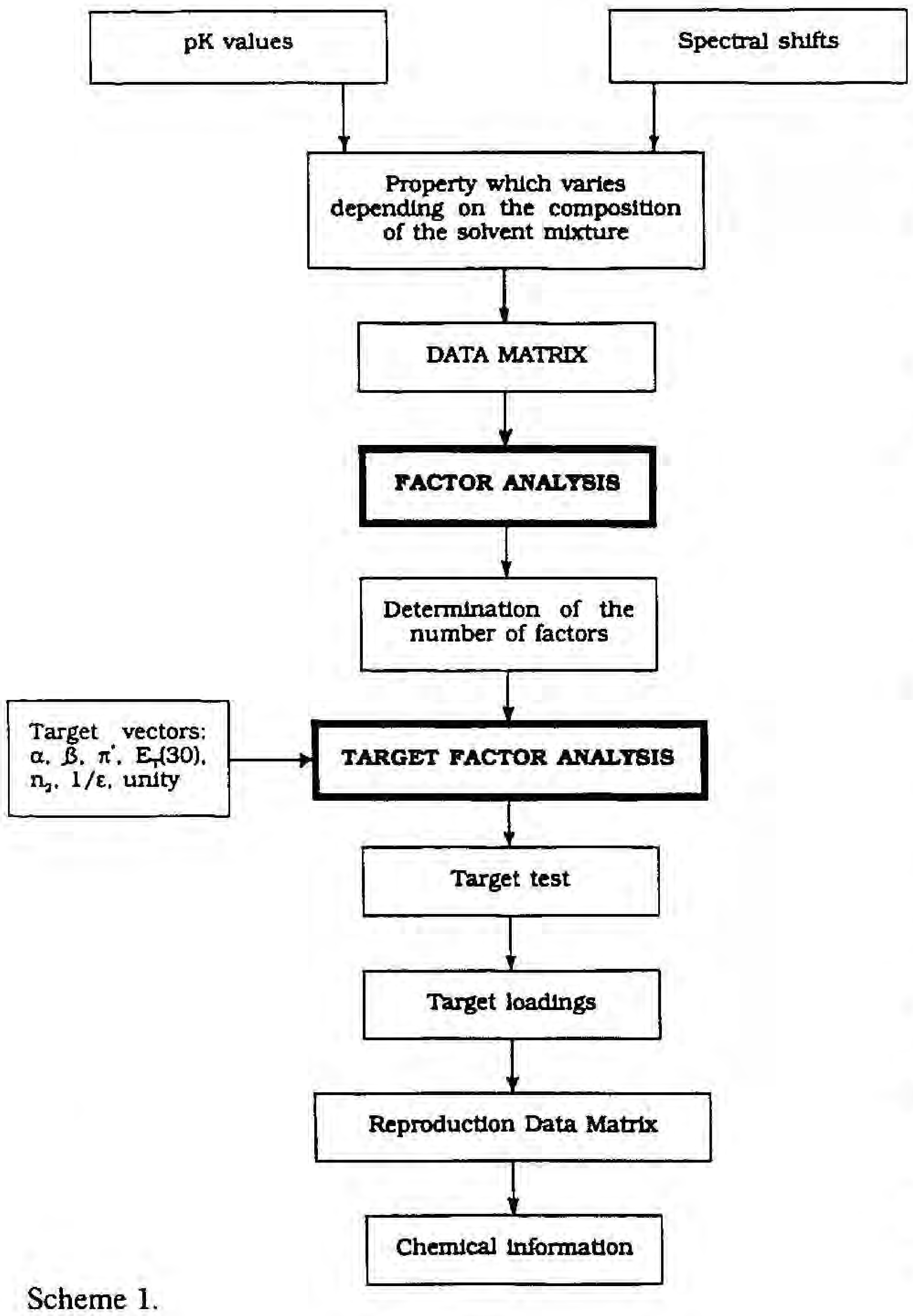
Correlation of solvent parameters with protonation constants in dioxane–water mixtures

Acid protonation constants for 3-hydroxynaphthalene-2-carboxylic acid (HNCA), propionic acid and salicylic acid were determined by the authors from e.m.f. readings [15]. The Gran method [16] was used for in-situ calibration of the cell and determination of the standard potential. After calibration, a known amount of acid solute was added to the cell and the titration continued until a suitable pH value was reached. In all the titrations, the titrand and the titrant were prepared in solvent mixtures of the same composition and at the same total concentration of inert electrolyte (0.1 *M* KNO_3 for 3-hydroxynaphthalene-2-carboxylic acid and 0.2 *M* KNO_3 for both salicylic and propionic acids). Experimental values were processed by the SUPERQUAD [17] program in order to determine the pK_a values. pK_a values for succinic acid were taken from the literature [18].

Data treatment

Different statistical tests, as proposed by Malinowski in the TARGET90 computer procedure [11], have been applied and are described briefly in Scheme 1.

Input data referring to the different properties, pK_a values or spectral shifts, which vary with the composition of the mixed solvent, are collected to build up the experimental data matrix **D**. The columns of the spectral shift data matrix are centred about the mean and normalized by dividing by their standard deviation (correlation about the mean). No pretreatment has been necessary in the case of the pK_a value data matrix. Factor analysis (FA) is used primarily to determine how many sources of independent variation (factors) are implied in the data. Two different tests are applied for this purpose, one derived from the application of the theory of error in FA as proposed by Malinowski [1], and the other one derived from the application of cross validation techniques [19,20]. Both the empirical indicator function, IND, which is calculated from the magnitude of the eigenvalues, and the standard error


pK values
Spectral shifts
Property which varies depending on the composition of the solvent mixture
DATA MATRIX
FACTOR ANALYSIS
Determination of the number of factors
Target vectors: α, β, π^*, $E_T(30)$, n_J, $1/\varepsilon$, unity
TARGET FACTOR ANALYSIS
Target test
Target loadings
Reproduction Data Matrix
Chemical information

Scheme 1.

of prediction, SEP, used in the second test, should have a minimum for the correct choice of factors.

Once the number of factors has been determined unambiguously, target factor analysis, TFA, is applied to identify the physical nature of these factors. The F statistical test as proposed by Malinowksi [11] has been used for this purpose; each of the targets is tested to be a possible candidate as a real factor: from the apparent error in the target, AET, and the error from the data matrix, EDM, the SPOIL value associated with the target is calculated [11,21]. Targets which lie in the factor space should have values of SPOIL below 3 and a percent significance level, %SL, above 5%. Those targets which are confirmed as possible real factors are used afterwards for the calculation of their target loadings and for the reproduction of the data matrix. The best combination of target test vectors which most accurately describes the data matrix is the one that yields the lowest error on the estimation of the loadings and the lowest root mean square error, RMS, in the reproduction of the data matrix. The best combination finally achieved is used to infer chemical information about the system and about the real causes affecting the properties of the solvent mixtures.

All the data treatment was performed using the TARGET90 program [11].

RESULTS AND DISCUSSION

Assessment of solvent parameters

Table 1 includes the data matrix. Results of eigenvalue analysis are shown in Fig. 1, where the evolution of IND (indicator function (1)) is plotted against the number of factors. Results of complete cross validation, plotted in Fig. 2, show the evolution of SEP (standard error of prediction functions) calculated according to Malinowski [19] or according to Wold [20]. A minimum at three factors can be seen for IND function and a common minimum at four factors for both Malinowski and Wold SEP functions. Since the RMS error obtained in the reproduction of the data matrix using three factors is quite acceptable, RMS = 0.041, this number of factors has been adopted for the target testing and for the combination of targets. Table 2 shows the results of target testing and the loadings associated with the targets taken as a definitive model.

Since the values of the longest-wavelength absorption maxima of the solvatochromic indicators have been preprocessed for correlation about the mean, the targets have been modified accordingly. This mathematical treatment has been used because of the big differences existing between the magnitude of the variations of the spectral shifts for indicator I and those for the other substances. In this way, a correction was made for the otherwise excessive statistical weight given to the data of betaine (indicator I) which would hide any other factor that did not directly depend on the

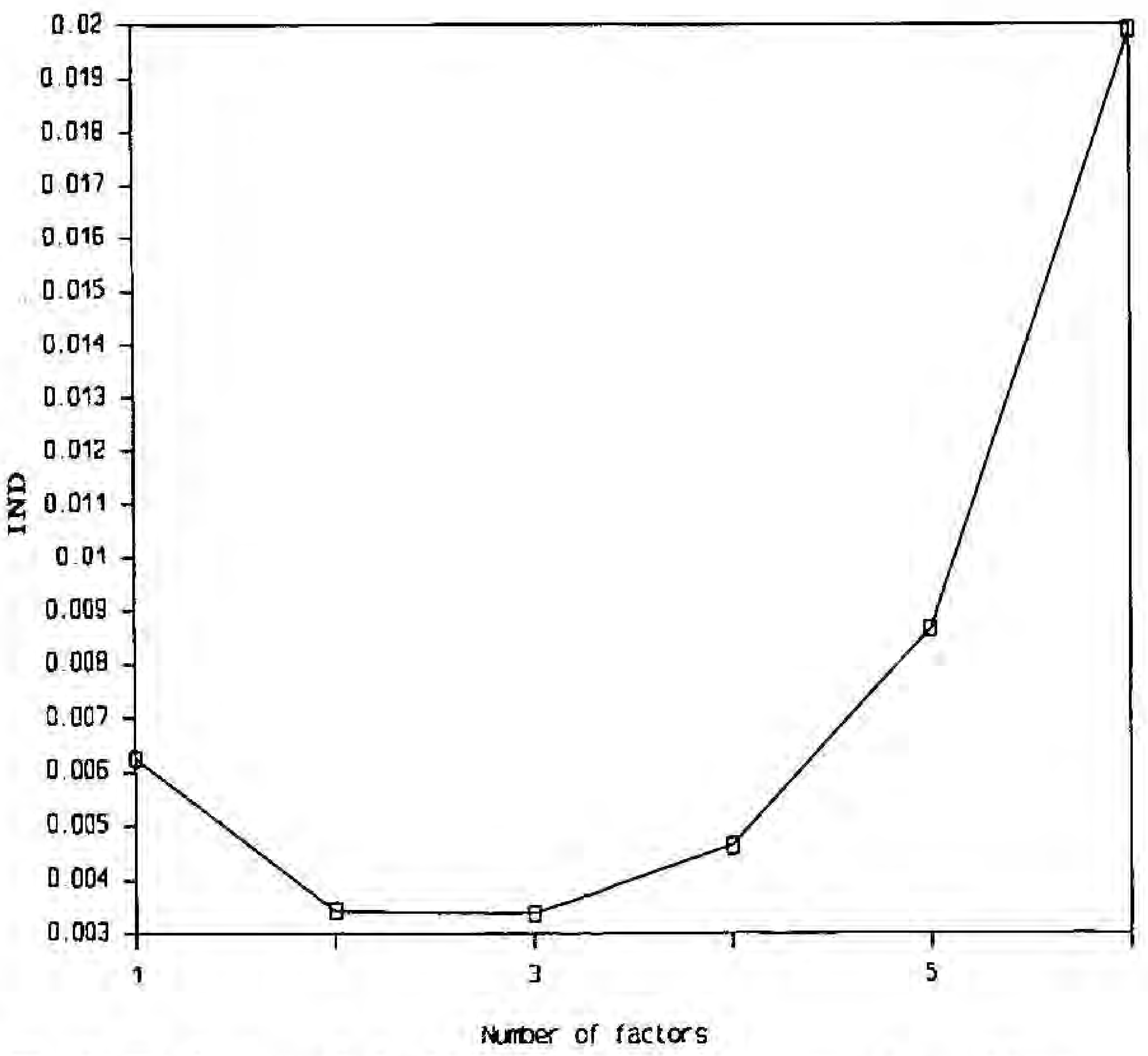


Fig. 1. Evolution of the IND function vs the number of factors.

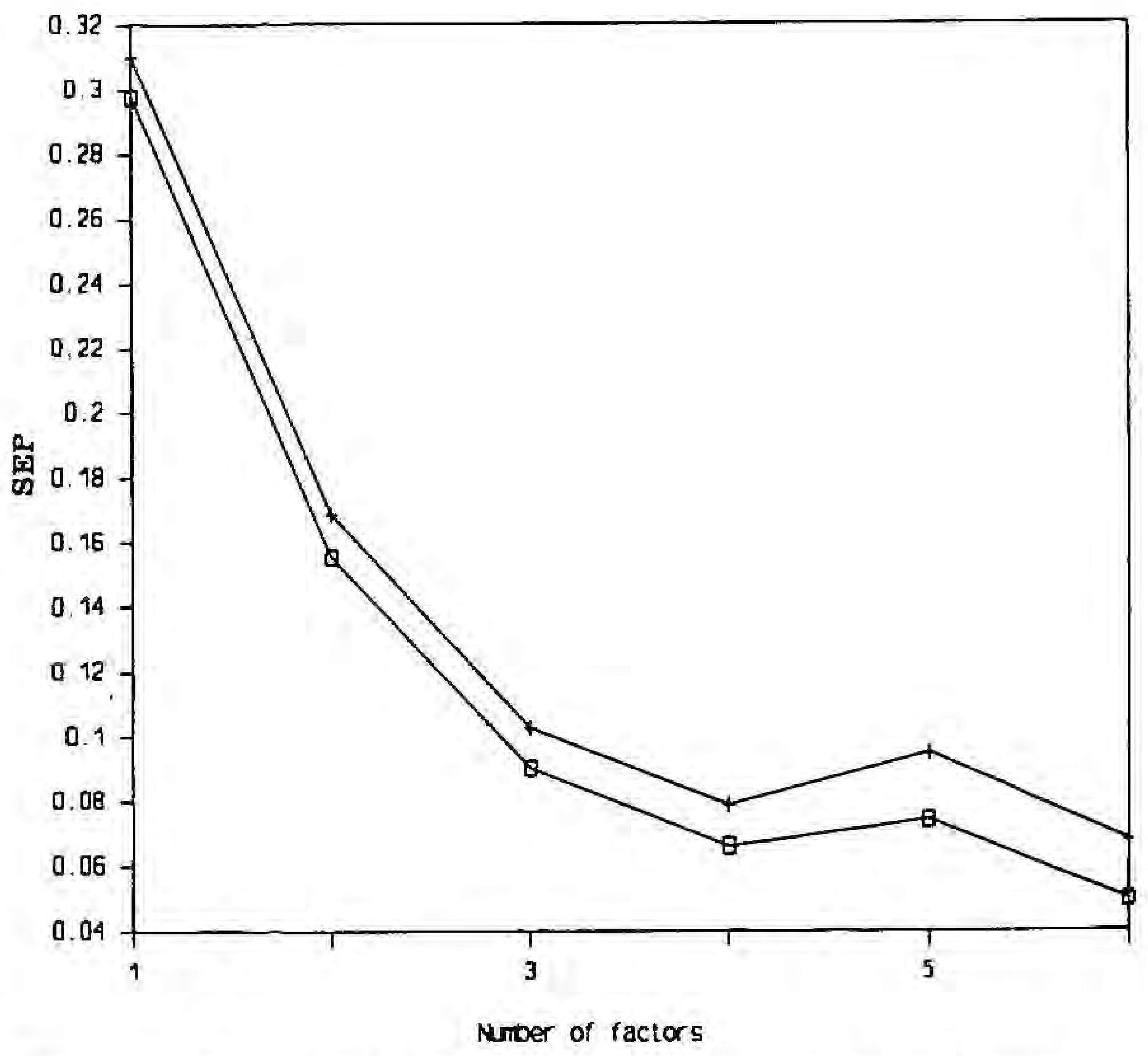


Fig. 2. Evolution of the SEP functions according to Malinowski (□) and according to Wold (+) vs. the number of factors.

shift of this substance. The normalization process enables to discover all the factors implied in the variation of the matrix data and does not cause any important loss of information because the origin of the targets tested is arbitrary.

The results of the eigenvalue analysis show that the variation of the absorption maxima of the compounds studied depends on three factors. Among all the sets of three targets which were combined, the one which best reproduced the original data matrix included α, β and π^* parameters, reflecting the hydrogen-bond solute–solvent interactions (α and β) and the influence of solvent polarity (π^*), as was expected from the model proposed by Kamlet and Taft.

TABLE 1

Data matrix. Frequency (in kK) of the longest-wavelength absorption maxima for several solvatochromic indicators in water–dioxane mixtures

I: 2,6-Diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenolate; II: 4-nitroanisole; III: 4-nitrophenol; IV: 4-ethylnitrobenzene; V: 2-nitroanisole; VI: *N*-methyl-2-nitroaniline; VII: 4-nitroaniline.

	Indicator						
	I	II	III	IV	V	VI	VII
10%	21.37	31.53(7) *	31.57(1)	35.06(1)	29.72(0)	22.41(0)	26.06(3)
20%	20.50	31.57(8)	31.51(4)	35.09(2)	29.76(1)	22.44(0)	26.05(2)
30%	19.97	31.61(9)	31.50(6)	35.21(2)	29.86(1)	22.53(1)	26.13(2)
40%	19.45	31.68(6)	31.49(3)	35.31(1)	29.97(2)	22.65(3)	26.26(2)
50%	18.75	31.84(6)	31.52(2)	35.43(4)	30.10(3)	22.78(2)	26.47(4)
60%	18.29	31.96(6)	31.58(5)	35.58(6)	30.29(2)	22.95(3)	26.52(5)
65%	18.00	32.03(3)	31.61(4)	35.67(7)	30.37(2)	23.01(5)	26.71(1)
70%	17.80	32.11(4)	31.64(3)	35.82(4)	30.44(0)	23.07(0)	26.82(6)
75%	17.57	32.15(5)	31.71(4)	35.81(8)	30.53(2)	23.14(0)	26.97(4)
80%	17.14	32.28(4)	31.77(1)	35.91(8)	30.61(0)	23.20(1)	27.06(2)
90%	16.33	32.43(3)	31.92(1)	36.15(3)	30.83(2)	23.32(5)	27.23(1)
95%	15.66	32.63(0)	32.10(6)	36.26(2)	31.00(2)	23.43(6)	27.46(2)
100%	12.59	32.86(2)	32.80(1)	36.46(4)	31.25(2)	23.61(7)	28.30(4)

* Values in parentheses are the associated errors with the last figure.

The experimental uncertainty of the frequency data can be expressed by the mean value of the individual standard deviations associated with each indicator in each water–dioxane mixture; this value is 0.032 kK.

Since the factor analysis for the working data matrix has been performed using normalized data, a comparison of the RMS error with the experimental uncertainty requires that this last magnitude be also normalized. The normalized experimental uncertainty is obtained as the average of the normalized standard deviations for each indicator by dividing the mean of the standard deviations associated with each experimental measurement for each solvent ratio (shown in parentheses in Table 1) by the same normalization factor used for the data matrix; that is, the standard deviation of the related frequency data (data columns of Table 1).

The associated RMS error in the reproduction of the data matrix is 0.049. This is a good description of the model since the normalized experimental uncertainty is equal to 0.078. The relatively large errors associated with some factor loadings can be explained because of the small importance of the contribution to which they are related.

The final loadings found for each solvatochromic indicator show the validity of all these compounds for the characterization of the solvent mixture studied. The criterion is that an indicator is really useful when it shows big loadings which are associated only with the causes of variation which are attributed to it. Thus, among the substances proposed as being sensitive to the change

TABLE 2

Target test and target loadings for spectral shift data

Target test

Number of factors = 3

Summary of target errors using 3 factors based on covariance.

Target	AET	EDM	SPOIL	F	df1	df2	%SL
n_2	1.20E − 01	3.86E − 02	2.94	3.95	10	4	9.9
$1/\epsilon$	1.15E − 01	5.06E − 02	2.04	2.12	10	4	24.4
α	3.86E − 02	8.10E − 02	0.00	0.09	10	4	99.9
β	1.08E − 01	1.01E − 01	0.38	0.47	10	4	84.9
π^*	1.94E − 02	2.76E − 02	0.00	0.20	10	4	98.2
E_TN	3.09E − 02	5.06E − 02	0.00	0.15	10	4	99.3
Unity	1.14E + 00	9.80E − 06	115694.40	99999.99	10	4	0.0

AET: Apparent error in the target (9)
EDM: Error from data matrix (9)
SPOIL: SPOIL associated with the target (9)
df1 and df2: Degrees of freedom (9)
%SL: Percent significance level (9)

Target loadings

Number of factors = 3

Factor loadings based on covariance.

	Alpha	Beta	π^*
I	0.65(2) *	0.029(6)	0.37(2)
II	0.06(5)	−0.07(2)	−1.08(5)
III	0.06(5)	−0.58(2)	−1.12(5)
IV	−0.00(6)	−0.03(2)	−0.98(6)
V	−0.03(2)	−0.00(8)	−0.97(2)
VI	−0.20(6)	0.12(2)	−0.76(6)
VII	−0.18(8)	−0.17(3)	−0.86(8)

* Values in parentheses are the errors associated with the last figure.

of polarity, a property which is well-defined by the π^* parameter, the indicators II, IV and V display a much larger value for the loading associated with π^* than for those related to α and β. This is a demonstration of the purity of their spectral shifts, which are related only to the solvent property that they measure. Indicator VI, by contrast, must be rejected as a polarity-measuring substance because of the important effect the hydrogen-bond interactions have on its spectral shift. These conclusions about the quality of this set of indicators agree with previous ones [22], reached using the graphic method proposed by Cheong and Carr [23]. Similar conclusions can be deduced for the rest of solvatochromic indicators.

The combination formed by $E_T(30)$ and β, which should offer the same chemical information as above, since $E_T(30)$ is a linear function of parameters π^* and α, yields a larger RMS error (0.154). This increase in error can be explained because of the fact that $E_T(30)$ does not allow the separation of the contributions that it contains, and these are not present simultaneously at the same ratio in all of the solvatochromic indicators used in this work.

Correlation of solvent parameters with protonation constants in dioxane–water mixtures

Table 3 includes the data matrix for protonation constants. Throughout the whole range of composition of the mixed solvent, each acidic compound has been studied in the presence of a constant concentration of inert electrolyte. Ionic strength, which affects the particular numerical value of a protonation constant in every given medium, does not seem to be a factor which modifies the kind of variation of pK_a with solvent composition in water–dioxane mixtures, since it is maintained constant over the full range of compositions studied and it is kept low enough to avoid spurious solvation influences. Actually, other authors have reached the same conclusions concerning the final correlations between pK_a and solvent-composition-dependent properties using very different working conditions (different inert electrolyte and different ionic strenghts). Although the protonation constants of the acidic substances used in this work have been determined at different values of ionic strength, this fact does not include the ionic strength as a new cause of variation of pK_a with solvent composition, but as a constant contribution to the pK_a value of each compound in the same way as the different chemical nature of each substance. The results of eigenvalue analysis are shown in Fig. 3, where the evolution of IND (indicator function) is plotted against the number of factors. The results of complete cross validation, plotted in Fig. 4, show the evolution of SEP (standard error of prediction) calculated according to Malinowski or according

TABLE 3

Data matrix. Acid dissociation constants (pK_a) of several carboxylic acids in water–dioxane mixtures

	Succinic1	Succinic2	Propionic	Salicylic	HNCA
10%	6.15	7.82	4.825(3) *	2.877(1)	[2.773] **
20%	6.57	8.27	5.091(6)	3.061(3)	2.964(2)
40%	7.04	8.99	5.676(1)	3.524(3)	3.448(7)
50%	7.34	9.36	6.063(1)	3.750(6)	3.762(4)
60%	7.71	9.92	6.494(1)	4.300(1)	[4.028]
70%	[8.30]	[10.72]	6.879(2)	4.655(1)	4.653(2)

* Values in parentheses are the associated errors with the last figure.

** Values in brackets are extrapolated.

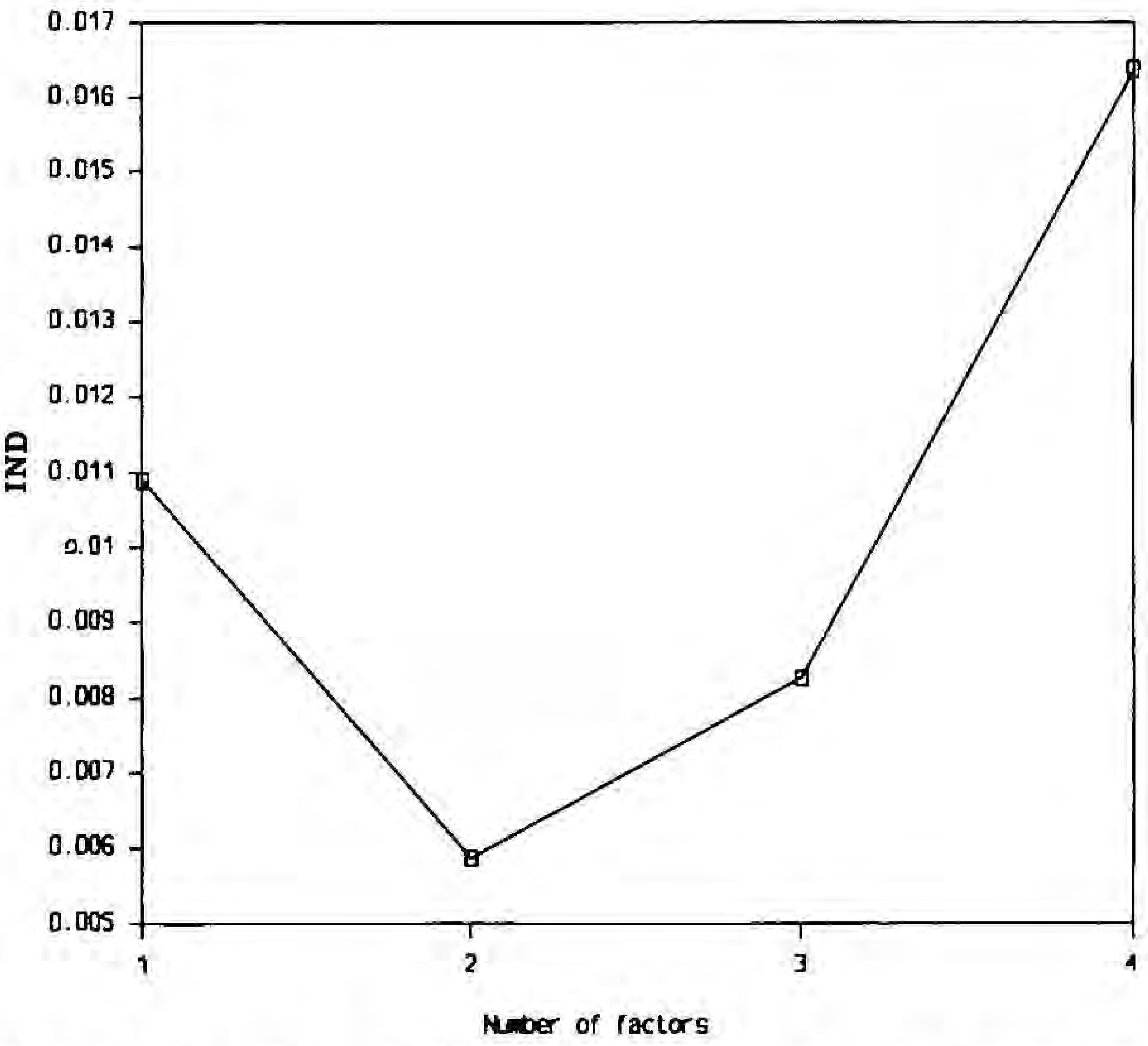


Fig. 3. Evolution of the IND function vs. the number of factors.

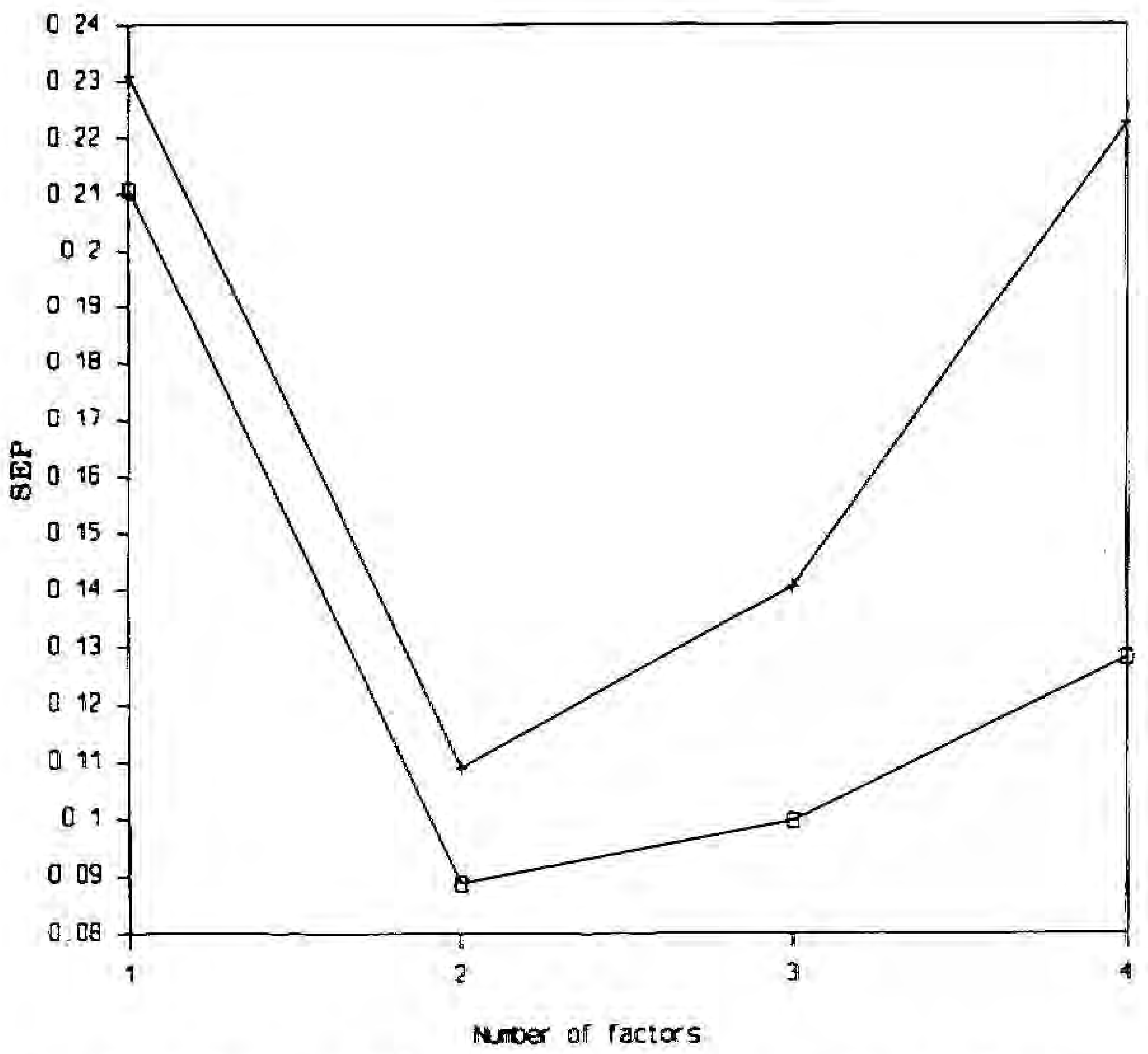


Fig. 4. Evolution of the SEP functions according to Malinowski (□) and according to Wold (+) vs. the number of factors.

to Wold. A common minimum is obtained for the three functions when the number of factors is equal to two. Thus, the variation of the pK_a values in water–dioxane mixtures, according to the results of factor analysis depends on two factors.

Table 4 shows the results of target testing and the loadings associated with the targets chosen as a definitive model. Among all the pairs of targets combined, a very good reproduction of the data matrix is obtained by combining π^* and unity as factors, with a RMS error = 0.086. Similar accuracy in the process of combination is reached using n_2 and unity as factors. These results do not justify the assumption that n_2 and π^* are equally valid in order to explain the variation of the pK_a values. It is a well known fact that the linear relationship between pK_a and n_2 [15] is broken when the amount of organic cosolvent is high. The

TABLE 4

Target test and target loadings for pK_a data

Target test

Number of factors = 2

Summary of target errors using 2 factors based on covariance.

Target	AET	EDM	SPOIL	F	dfl	df2	%SL
π^*	3.36E − 02	3.44E − 02	0.00	0.32	4	3	85.1
α	8.72E − 02	3.24E − 02	2.50	2.41	4	3	24.8
β	1.51E − 02	2.15E − 03	6.99	16.55	4	3	2.2
$1/\epsilon$	3.33E − 03	1.45E − 03	2.06	1.75	4	3	33.8
n_2	1.22E − 02	1.36E − 02	0.00	0.27	4	3	88.3
$E_T(30)$	2.52E + 00	1.53E + 00	1.31	0.90	4	3	55.7
Unity	2.39E − 02	1.82E − 02	0.85	0.57	4	3	70.5

AET: Apparent error in the target (9)

EDM: Error from data matrix (9)

SPOIL: SPOIL associated with the target (9)

dfl and df2: Degrees of freedom (9)

%SL: Percent significance level (9)

Target loadings

Number of factors = 2

Factor loadings based on covariance.

	π^*	Unity
Succinic1	− 6.6(4) *	13.9(4) *
Succinic2	− 9.2(4)	18.5(5)
Propionic	− 6.9(3)	12.8(3)
Salicylic	− 6.0(2)	9.8(2)
HNCA	− 6.0(3)	9.7(3)

* Values in parentheses are the associated errors with the last figure.

lack of pK_a data from the literature for dioxane–water mixtures at high cosolvent concentrations does not allow extension of the matrix to show the true difference between the two pairs of targets compared. The correct reproduction of the data matrix using the target π^* and unity implies that the Kamlet and Taft general equation is reduced in this case to only two terms to describe the modification of the property studied in dioxane–water mixtures: the first one, derived from the target unity, represents the value of the acid dissociation constant of the different solutes in a hypothetical solvent with $\alpha = \beta = \pi^* = 0$, and the second one is related to the polarity–polarizability of the solvent, which is well characterized by the π^* parameter.

Since the solvent polarity is identified as the main cause of variation of the pK_a values in the dioxane–water mixtures, it seems surprising that other parameters proposed as a measure of polarity, such as $E_T(30)$ and $1/\epsilon$, yield much larger errors in the reproduction of the original data matrix. When the combination of $E_T(30)$ and unity is tested, there is an increase in the RMS error to 0.165. This unexpected behaviour can be attributed to the 'polluted' nature of the $E_T(30)$ parameter, which contains an important contribution from the hydrogen bond interaction, which leads to a significantly worse description of the system.

On the other hand, the bad results obtained from the combination of $1/\epsilon$ and unity, with an RMS error = 0.174, confirm that the bulk properties of the solvent (macroscopic properties) are very different from those related with the cybotactic zone (microscopic properties). Only a knowledge of these microscopic properties reflected by the solvatochromic parameters will provide a correct description of microscopic processes, such as acid–base equilibria.

REFERENCES

1 E.R. Malinowski and D.G. Howery, *Factor Analysis in Chemistry*, Wiley, New York, 1980.

2 W.R. Fawcett and T.M. Krygowski, A characteristic vector analysis of solvent effects for thermodynamic data, *Canadian Journal of Chemistry*, 54 (1976) 3283–3292.

3 J.T. Edward and Sin Cheong Wong, Ionization of carbonyl compounds in sulphuric acid. Correction for medium effects by characteristic vector analysis, *Journal of the American Chemical Society*, 99 (1977) 4229–4232.

4 J.R. Haak and J.B.F.N. Engberts, Solvent polarity and solvation effects in highly aqueous mixed solvents, Application of the Dimroth–Reichardt $E_T(30)$ parameter. *Recueil de Travaux en Chimie des Pays Bas*, 105 (1986) 307–311.

5 E. Grundwald and S. Winstein, The correlation of solvolysis rate, *Journal of the American Chemical Society*, 70 (1948) 846–854.

6 E.M. Kosower, *An Introduction to Physical Organic Chemistry*, Wiley, New York, 1968.

7 C. Reichardt, *Solvents and Solvent Effects in Organic Chemistry*, VCH, Weinheim, 1988.

8 M.J. Kamlet, J.L.M. Abboud and R.W. Taft, An examination of linear solvation energy relationships, in R.W. Taft (Editor), *Progress in Physical Organic Chemistry*, Vol. 13, Interscience, New York, 1981, pp. 445–630.

9 P. Suppan, Local polarity of solvent mixtures in the field of electronically excited molecules and exciplexes, *Journal of the Chemical Society Faraday Transactions 1*, 83 (1987) 495–509.

10 H. Langhals, Eine quantitative Beschreibung der Solvatochromie in binären Flüssigkeitsgemischen, *Nouvelle Journal de Chimie*, 6(6) (1981) 97–99.

11 E.R. Malinowski, Statistical *F*-test for abstract factor analysis and target testing, *Journal of Chemometrics*, 3 (1988) 49–60.

12 A.I. Vogel, *A Textbook of Practical Organic Chemistry*, Longmans Green, London, 5th ed., 1989, p. 407.

13 J.E. Powell and M.A. Hiller, Preparation of carbonate-free bases, *Journal of Chemical Education*, 34 (1957) 330.

14 H. Langhals, Polarity of binary liquid mixtures, *Angewandte Chemie International Edition in English*, 21 (1982) 724–733.

15 E. Casassas and G. Fonrodona, Protonation equilibria of 3-hydroxy-2-naphthoic acid in dioxane–water solution: effect of the solvent composition and of the ionic strength of the medium, *Journal de Chimie Physique*, 86 (1989) 391–402.

16 G. Gran, Determination of equivalent point in potentiometric titrations, *Acta Chemica Scandinavica*, 4 (1950) 559–577; Determination of the equivalence point in potentiometric titrations, II, *The Analyst (London)*, 77 (1952) 661–671.

17 P. Gans, A. Sabatini and A. Vacca, SUPERQUAD: an improved general program for computation of formation constants from potentiometric data, *Journal of the Chemical Society, Dalton Transactions*, (1985) 1195–1200.

18 M. Yasuda, Dissociation constants of some carboxylic acids in mixed aqueous solvents, *Bulletin of the Chemical Society Japan*, 32 (1959) 429–432.

19 E.R. Malinowski, Theory of the distribution of error eigenvalues resulting from principal component analysis with applications to spectroscopic data, *Journal of Chemometrics*, 1 (1987) 33–40.

20 S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, *Technometrics*, 20 (1978) 397–405.

21 E.R. Malinowski, Theory of the error for target factor analysis with applications to mass spectrometry and nuclear magnetic resonance spectrometry, *Analytica Chimica Acta*, 103 (1978) 339–354.

22 E. Casassas, G. Fonrodona and A. de Juan, Determinación del parámetro de polaridad–polarizabilidad π^* y correlación de éste con $E_T(30)$ para mezclas dioxano–agua, Anales de Quimica, 87 (1991), in press.

23 W.J. Cheong and P.W. Carr, Kamlet–Taft π^* polarizability/dipolarity of mixtures of water with various organic solvents, *Analytical Chemistry*, 60 (1988) 820–826.

Analytica Chimica Acta, 283 (1993) 548–558
Elsevier Science Publishers B.V., Amsterdam

Factor analysis applied to the study of the effects of solvent composition and nature of the inert electrolyte on the protonation constants in dioxane–water mixtures

E. Casassas, N. Domínguez, G. Fonrodona and A. de Juan

Departament de Química Analítica, Universitat de Barcelona, Avda. Diagonal 647, 08028 Barcelona (Spain)

(Received 10th September 1992; revised manuscript received 8th December 1992)

Abstract

The effect of solvent properties on protonation constants was studied for water–dioxane mixtures covering a wide range of solvent compositions, from data sets on several solutes with a variety of functional groups. The establishment of the model was carried out using different approaches: the use of a classical procedure, where the best multiparametric fit to the Kamlet and Taft equation was evaluated for each substance; and the application of combined factor analysis and target factor analysis to quantify and identify the factors affecting the variation of the whole data sets, without the need to postulate any a priori hypothetical model. Further, a preliminary overview of the influence of different inert electrolytes on the values of the solute protonation constants in this solvent mixture was also obtained through factor analysis.

Keywords: Dioxane; Factor analysis; Protonation constants; Solvatochromic parameters; Solvent effects; Waters

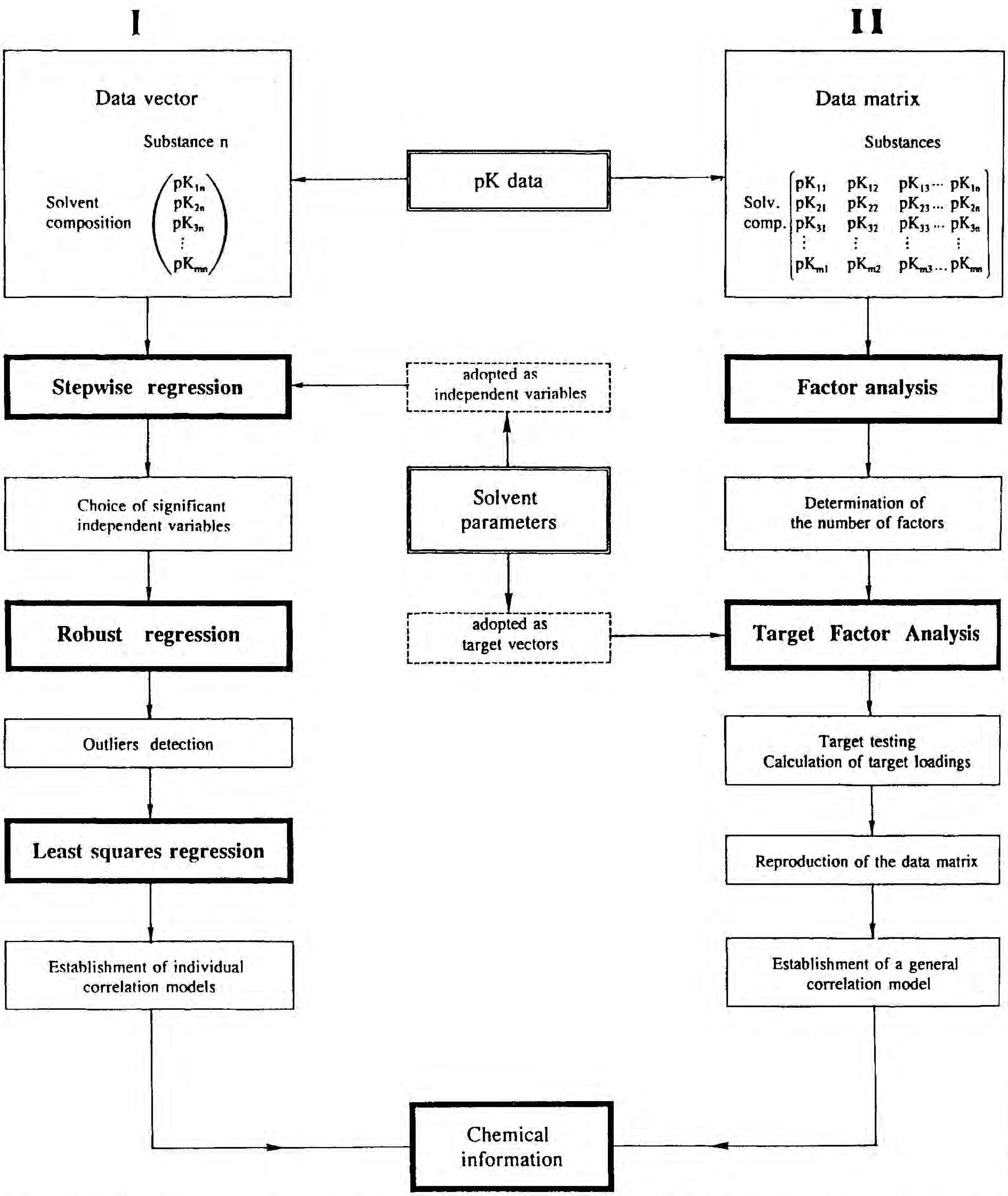
Factor analysis (FA) has been demonstrated to be a powerful technique in the study of complex chemical phenomena, whose interpretation requires multivariate approaches [1]. It has often been used to explain solute–solvent interactions, as a helpful tool in the field of linear free energy relationships [2] when medium effects have to be described and to understand better some solvent effects on solute physico-chemical properties [3–5]. This wide application in the study of pure solvent effects contrasts with the scarceness of analogous works devoted to solvent mixtures.

In the latter field, in previous studies [6] FA was used to find the number of factors involved in the variation of the protonation constants of carboxylate groups of several substances when the composition of water–dioxane mixtures is modified and target factor analysis (TFA) was then applied to identify these factors. The analysed data set revealed that this variation could be expressed through the general equation $\mathrm{p}K_s = \mathrm{p}K_0 + s\pi^*$, where π^* is the microscopic polarity parameter proposed by Kamlet et al. [7], and s is a constant.

In this work, the same study was extended to several acidic solutes with different functional groups (phenolic and amino groups) and to a wider composition range of mixtures, in order to propose a more general model. To achieve this objective, two different parallel approaches were applied, as described briefly in Scheme 1.

The first attempt was carried out in a classical way, trying to fit the behaviour of each substance to a previously proposed model, using the more extended form of the multiparametric equation proposed by Kamlet et al. [7], $XYZ = XYZ_0 + a\alpha$

Correspondence to: E. Casassas, Departament de Química Analítica, Universitat de Barcelona, Avda. Diagonal 647, 08028 Barcelona (Spain)



Scheme 1. Outline of the approaches used: classical approach (data treatment I) and FA + TFA approach (data treatment II).

$+b\beta+s\pi^*$, where α, β and π^* are solvatochromic parameters related to the hydrogen bond acidity, hydrogen bond basicity and polarity–polarizability, respectively [7–9], *XYZ* is the solute property studied (pK_s) and XYZ_0 is the hypothetical value for this property in a solvent where $\alpha=\beta=\pi^*=0$. For this purpose, a stepwise procedure was first applied to select the significant solvent properties to be included in the model. Thereafter, a robust analysis was performed to detect outliers, and finally a least-squares fit was carried out with the appropriate data and variables to lead to the definite expressions for each compound. The second approach involved the application of the hypothesis-free model technique, FA, to establish how many sources of variation affect the solute property studied. Thereafter, these factors were identified using TFA and combined to yield the best general model for all the substances studied.

The classical approach gives force-fitted equations, whereas FA provides a much less arbitrary criterion for selecting the quantity and nature of the possible causes of variation in the data sets, avoiding the introduction of a priori hypotheses. From the combination of the two procedures, interesting conclusions can be inferred relating to the variation of protonation constants with the solvent composition in water–dioxane mixtures.

To complete the study of the acid–base behaviour of solutes in water–dioxane mixtures, a preliminary overview of the influence of different inert electrolytes on the values of the solute protonation constants in this solvent system by FA was also obtained.

EXPERIMENTAL

Reagents

Analytical-reagent grade chemicals were used, unless indicated otherwise. Water was deionized and distilled twice over potassium permanganate.

Dioxane (Probus) was purified by Eigenberger's method [10]. 4-Nitroanisole (Merck) was purified by treatment with active carbon in an acetone solution and recrystallization in water. 2-Nitroanisole, 4-ethylnitrobenzene and *N*,*N*-diethyl-4-nitroaniline were obtained from Aldrich, 4-nitroaniline and sodium salicylate from Carlo Erba and 2,6-diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenolate from Professor E. Bosch. Potassium hydroxide (Merck) was prepared as a CO_2-free solution in dioxane–water by the ion-exchange procedure [11]. Sodium hydroxide, nitric acid, propionic acid, potassium nitrate, sodium nitrate and potassium chloride were obtained from Merck. Salicylic acid (Merck) was purified by sublimation and 3-hydroxynaphthalene-2-carboxylic acid (Merck) by treatment with active carbon and recrystallization from ethanol.

Apparatus

A Beckman DU-7 spectrophotometer was interfaced (RS232) to an IBM personal computer. Acquisition of spectra was controlled through Beckman data capture software. An Orion SA 720 potentiometer (precision ±0.1 mV), an Orion 90-05 AgCl/Ag reference electrode with ceramic junction and internal reference solution of saturated KCl in the working aqueous–organic mixture, an Orion 91-01 glass electrode, a double-jacketed cell thermostated at 25 ± 0.1°C, a Metrohm 665 Dosimat autoburette (precision 0.01 ml) with an exchange unit of 5 cm^3 with an antidiffusion burette tip and a magnetic stirrer were used. The whole titration set-up was connected to a personal computer (HP Vectra ES/12 or HP 9133) through an HP 3421 A interface, which allows full automation of the titration process.

Determination of acid dissociation constants (pK_a) of several acidic solutes in water–dioxane mixtures

pK_s values for 3-hydroxynaphthalene-2-carboxylic acid (HNCA), propionic acid and salicylic acid using KNO_3 as inert electrolyte were determined previously [12,13]. Values for propionic acid in other inert electrolytes were obtained by potentiometric titrations performed at a constant temperature (25°C) and constant ionic strength. The potentiometric cell used was GE/working solution (b mol l^{-1})–n% dioxane/RE (KCl$_{sat}$, n% dioxane), where GE is the glass electrode, RE the reference electrode, b the constant concentration of inert electrolyte and *n* the percentage of dioxane in the working aqueous–organic mixture. The Gran method [14] was used for in situ calibration of the cell and determination of the standard potential. After calibration, a known amount of acid solute was added to the cell and the titration continued until a suitable pH value. The experimental e.m.f. readings were processed with the SUPERQUAD program [15] to obtain the pK_s values. Glycine pK_s values were taken from the literature [16]. All the pK_s values and the experimental conditions used in their determination are given in Table 1.

Determination of solvatochromic parameters in water–dioxane mixtures

Solvatochromic parameters were determined from spectroscopic measurements of certain substances, the so-called solvatochromic indicators, whose shift in the wavelength of maximum absorption is sensitive only to the solvent property that must be measured and is free from contributions produced by other causes. For amphiprotic solvents or amphiprotic mixtures, the solvatochromic indicators proposed by Kamlet and Taft were used as follows: 2-nitroanisole, 4-nitroanisole and 4-ethylnitrobenzene (all of them showing spectral shifts sensitive only to the polarity of the solvent) were used to determine the π^* parameter; both 2,6-diphenyl-4-(2,4,6-triphenyl-1-pyridinio)phenolate (whose spectra shifts are sensitive to the polarity and to the hydrogen bond donor acidity of solvent) and 4-nitroanisole were used to determine the α parameter and both 4-nitroaniline (with spectral shifts sensitive to the polarity and to the hydrogen bond acceptor basicity of solvent) and *N,N*-diethyl-4-nitroaniline (whose spectral shifts are sensitive only to the polarity) were used to determine the β parameter. All the wavelength values used to evaluate the solvatochromic parameters were determined by the authors [12,13,17], except several betaine wavelengths values which were taken from the literature [18].

To determine the required wavelengths, the spectrum of each test solution (solvatochromic indicator in the binary mixture) is recorded against a blank formed by a solvent mixture of composition identical with that of the solvent used in the test solution. Both the test solution and the blank circulate through continuous-flow systems in which the mixture composition is changed stepwise with additions of dioxane until the whole composition range of the binary mixture is covered. Replicates of each spectrum are made from identical independently prepared test solutions. Thereafter, a digitization procedure yields the wavelength corresponding to the ab-

TABLE 1

Experimental conditions and acid dissociation constants (pK_a) of several acidic compounds in water–dioxane mixtures

Experimental conditions							
Substance	Propionic acid	Propionic acid	Salicylic acid [a]	Salicylic acid [b]	HNCA [a,c]	Glycine [a]	Glycine [d]
Ionic strength (M)	0.2	0.2	0.2	0.2	0.1	0.1	0.1
Inert electrolyte	KNO_3	$NaNO_3$	KNO_3	KNO_3	KNO_3	$NaClO_4$	$NaClO_4$
Dioxane (%, v/v)	pK_a values						
10	4.825	4.8334	2.897	12.789			
20	5.091		3.067	13.23	2.964	2.6	9.64
30	5.395	5.382	3.318	13.43	3.171		
40	5.676		3.524	13.51	3.448	2.94	9.7
50	6.063	6.0236	3.789	13.94	3.762	3.17	9.76
60	6.494		4.305	14.12		3.45	9.84
65	6.745		4.478	14.46	4.382		
70	6.879	6.861	4.654	14.84	4.653	3.81	9.98
75		7.052			4.998	4.05	10.07
80					5.289	4.33	10.23

[a] COOH group. [b] OH group. [c] HNCA = 3-hydroxynaphthalene-2-carboxylic acid. [d] NH_2 group.

TABLE 2

Solvent parameters for water–dioxane mixtures

Dioxane (%, v/v)	α	β	π^*	n_2
10	0.94	0.217	1.143	0.023
20	0.81	0.296	1.124	0.05
30	0.74	0.379	1.088	0.083
40	0.67	0.433	1.049	0.123
50	0.61	0.474	0.989	0.174
60	0.57	0.486	0.92	0.241
65	0.55	0.465	0.885	0.282
70	0.54	0.469	0.849	0.33
75	0.51	0.501	0.823	0.388
80	0.48	0.532	0.781	0.458
85	0.45	0.552	0.748	0.545
90	0.4	0.551	0.704	0.655
95	0.35	0.518	0.636	0.801
100	−0.07	0.360	0.54	1

sorption maximum. The gross averages of the different wavelengths obtained for each solvent composition conveniently processed give the values of the solvatochromic parameters. The β values have been revised and modified in this work. Numerical values for α, β and π^* parameters are given with some other solvent parameters in Table 2.

Study of the variation of protonation constants in water–dioxane mixtures

Classical approach. For each substance, a correlation model based on the general equation of Kamlet et al. [7], $XYZ = XYZ_0 + a\alpha + b\beta + s\pi^*$, where XYZ is the property of the solvent studied (in this instance pK_a) and α, β and π^* are the microscopic parameters related to hydrogen bond donor acidity, hydrogen bond acceptor basicity and polarity–polarizability of the solvent, respectively, is proposed. As was pointed out by Kamlet and Taft [19], the number of terms in the equation can be larger or smaller depending on the significance of the different solvent parameters with respect to the solute property being studied.

The effect of the different solvent properties (quantified through solvatochromic parameters) on the variation of pK_s values of one substance over the entire range of mixture compositions (dependent variable) is first analysed using a stepwise procedure, in which α, β and π^* are the possible independent variables, whose significance is tested. The stepwise procedure used is a combination of both forward selection and backward elimination of variables, as implemented in the PARVUS statistical package [20]. For each selection step of a new independent variable the

TABLE 3

A.1 and A.2 data matrices [a]: pK_a values for several acidic compounds in water–dioxane mixtures of different composition

Dioxane (%, v/v)	Compound [b]					
	HNCA	gly1	gly2	PropK	Sal1	Sal2
20	2.964	2.6	9.64	5.091	3.067	13.23
30	3.171	2.71	9.67	5.395	3.318	13.43
40	3.448	2.94	9.7	5.676	3.524	13.51
50	3.762	3.17	9.76	6.063	3.789	13.94
60	4.288	3.45	9.84	6.494	4.305	14.12
70	4.653	3.81	9.98	6.879	4.654	14.84
75	4.998	4.05	10.07			
80	5.289	4.33	10.23			

[a] Dashed box = A.1 data matrix; solid box = A.2 data matrix. [b] HNCA = 3-hydroxynaphthalene-2-carboxylic acid (COOH group); gly1 = glycine (COOH group); gly2 = glycine (NH_2 group); PropK = propionic acid at ionic strength 0.2MKNO$_3$; Sal1 = salicylic acid (COOH group); sal2 = salicylic acid (OH group).

program computes alternatively one of the following F-tests. In the first, an F-to-enter value is computed for each non-entered variable to evaluate the significance of the decrease in variance in the fit. In the second F-test, an F-to-delete value is computed for the variables already entered. The process continues until any one of the unselected variables gives F values larger than the user-defined control F value. Once it has been decided which of the variables in the model are significant, the regression analysis is performed by using the PROGRESS program [21]. The first step in this program includes a robust approach [least median of squares regression (LMS)] to detect the presence of outliers. The outliers diagnostic is performed through the application of two different high breakdown point methods, such as the study of the standardized LMS residuals and the robust diagnostic method proposed by Rousseeuw and Leroy [21]. To avoid the removal of good leverage data points, only data to be considered as outliers through both previous methods have been removed by the authors. Thereafter, a least-squares fit is performed to obtain the definite model for each substance.

FA and TFA approach. Input data referring to pK_a for all substances and all mixture compositions are collected to build up two different experimental data matrices. In the first (A.1 data matrix), the main interest is in finding a general model for substances with different functional groups; in the second (A.2 data matrix), the general model is deduced from pK_a data for a smaller number of solutes, but covering a wider range of mixture compositions. This separate treatment was adopted to avoid the use of extrapolated pK_s data, which necessarily would come from the application of an external hypothetical chemical model. For both matrices, no pretreatment of the data was necessary. The A.1 and A.2 data matrices are shown in Table 3.

Factor analysis (FA), also known as principal component analysis (PCA), has been primarily used to determine how many causes affect the variation of pK_a values with solvent composition. The correct number of abstract factors is obtained through (1) the application of the IND function [22], (2) the study of the evolution of standard error in prediction (SEP) function, through cross-validation techniques, as proposed by Malinowski [22] and Wold [23], (3) the application of an F-test to check the significance of the eigenvalue [24] and (4) the study of the real error (RE) [22]. Both the empirical indicator function, IND (which is calculated from the magnitude of the eigenvalues), and standard error of prediction, SEP, should have a minimum for the correct number of factors (NF). When the F-test is performed, the number of factors is obtained from the break in the evolution of the values of the percentage of significance level (%SL, Table 6) associated with the successive F values evaluated for an increasing number of factors. The correct NF is given by the RE function when the difference between the calculated value of this function and the experimental error passes through a minimum. These tests yield the number of factors which explain the variation of data; they are abstract factors and the assignment of a physical meaning to them requires further information and data treatment.

For the A.1 data matrix all four methods were applied, whereas for the A.2 data matrix only methods 3 and 4 were used.

When the number of factors has been established, TFA is developed to identify the chemical nature of these factors. Up to this point, in contrast with the classical approach, no hypothetical model is necessary. A target is defined as a vector representing known experimental data other than those under analysis, which is tentatively proposed as a possible real factor causing the observed variation of the data within the original data matrix. Geometrically, a target vector accepted as a real factor lies in the space determined by the abstract factors found in FA.

As both the A.1 and A.2 data matrices are formed by equilibrium constants determined at different solvent compositions, the variation within the original data matrices has to be related to modifications in the solute–solvent interactions, well described by the Kamlet and Taft parameters. Hence these solvent parameters and the mole fraction are accepted as targets and checked as possible real factors present in a future general model. No other solvent parameters

TABLE 4

B.1 and B.2 data matrices [a]: pK_a values for several acidic compounds in water–dioxane mixtures of different composition

Compound [b]	Dioxane (%, v/v)				
	20	40	50	70	60
HCNA	2.964	3.448	3.762	4.653	4.288
Sal1	3.067	3.524	3.789	4.654	4.305
Sal2	13.23	13.51	13.94	14.84	14.12
PropK	5.091	5.676	6.063	6.879	6.494
gly1	2.6	2.94	3.17	3.81	3.45
gly2	9.64	9.7	9.76	9.98	9.84

[a] Dashed box = B.1 data matrix; solid box = B.2 data matrix.
[b] See Table 3.

(macroscopic or microscopic) have been proposed as possible targets, because previous studies yielded much worse results for them in comparison with those related to the chosen parameters [6]. The validity of each target is analysed through an F-test [24] and evaluating its SPOIL function [25]. Targets which lie in the factor space should have SPOIL values below 3 and a percent significance level, %SL, for the F-test above 5. The accepted targets are thereafter combined to reproduce the original data matrix. The best combination of target test vectors is that which yields the lowest error in the estimation of the loadings and the lowest root mean square error (RMS) in the reproduction of the data matrix.

All the data treatment was performed using the TARGET90 program [24].

Study of the effect of the inert electrolyte on the determination of protonation constants in water–dioxane mixtures

Because it has not been possible to accept any general model that could explain the effect of using different inert electrolytes on the values of protonation constants in water–dioxane mixtures, no classical procedure of model fitting was used for data treatment.

FA was then applied to detect the possible existence of the effect referred to above, by comparison of the number of factors calculated for the two following matrices: (B.1) pK_s data for several solutes determined at the same solvent composition and using the same inert electrolyte (KNO_3); and (B.2) the B.1 data matrix enlarged

TABLE 5

Individual pK_a correlation models, according to the classical approach

Solute [a]	Significant variable [b]	Correlation model [c,d]
Propionic acid,	α, π^*	p$K = -1.3(2)\alpha - 5.4(3)\pi^* + 12.1(1)$
$I = 0.2$ M KNO_3		$r^2 = 0.999$; sc.est. [e] = 0.03
Propionic acid,	π^*	p$K = -6.7(3)\pi^* + 12.5(3)$
$I = 0.2$ M $NaNO_3$		$r^2 = 0.994$; sc.est. = 0.08
Sal1,	π^*	p$K = -5.9(1)\pi^* + 9.7(1)$
$I = 0.2$ M KNO_3		$r^2 = 0.996$; sc.est. = 0.04
Sal2,	π^*	p$K = -5.9(5)\pi^* + 19.7(5)$
$I = 0.2$ M KNO_3		$r^2 = 0.960$; sc.est. = 0.1
HNCA,	π^*	p$K = -6.6(2)\pi^* + 10.4(2)$
$I = 0.1$ M KNO_3		$r^2 = 0.991$; sc.est. = 0.09
Gly1,	π^*	p$K = -4.9(3)\pi^* + 8.0(2)$
$I = 0.1$ M $NaClO_4$		$r^2 = 0.980$; sc.est. = 0.08
Gly2,	π^*	p$K = -1.4(1)\pi^* + 11.2(1)$
$I = 0.1$ M $NaClO_4$		$r^2 = 0.950$; sc.est. = 0.04

[a] For abbreviations, see Table 3. [b] Significant variables were deduced through stepwise regression (SR). [c] Correlation models were obtained from the successive application of robust regression [least median of squares regression (LMS)] and least-squares regression without the presence of outlier data. [d] Values in parentheses are the errors associated with the last figure. [e] Scale estimate.

by the inclusion of pK_s data for both glycine functional groups, determined employing $NaClO_4$ as an inert electrolyte.

An increase in the number of factors from B.1 to B.2 could probably be attributed to the effect of the new different electrolyte. The determination of the correct number of factors was carried out with the four methods listed previously. The small size of the B.1 data matrix limited the range of procedures to the study of RE and the significance of eigenvalues, whereas with the B.2 data matrix all methods were applied. Both B.1 and B.2 data matrices are given in Table 4.

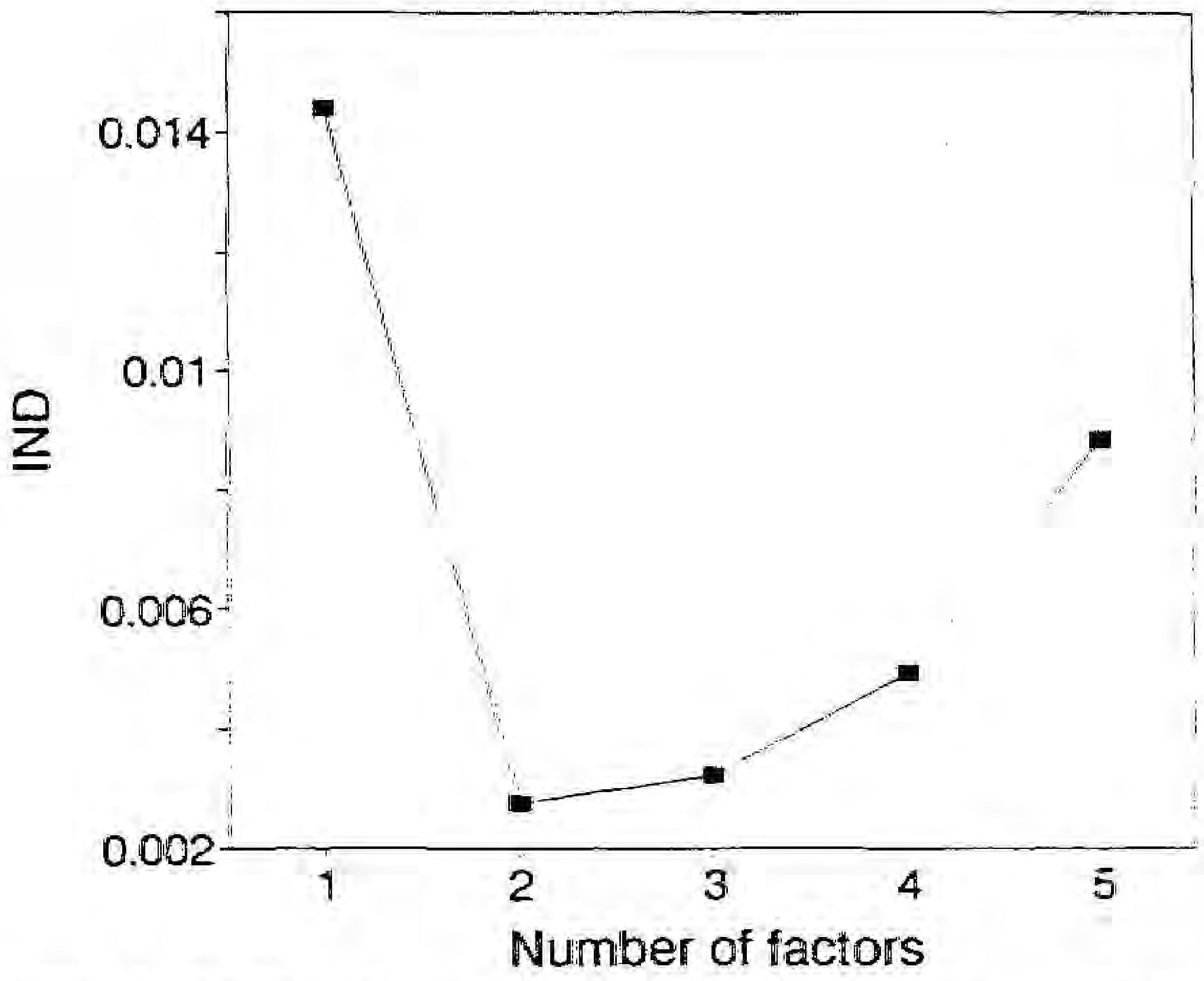


Fig. 1. Evolution of the IND function vs. the number of factors for the A.1 data matrix.

RESULTS AND DISCUSSION

Study of the variation of protonation constants in dioxane–water mixtures

Table 5 gives the results provided by the classical approach. For each substance, the significant independent variables selected by applying the stepwise procedure to build up the model are listed, together with the definitive fit to the appropriate form of the Kamlet and Taft equation with its associated quality parameters.

The results obtained with the application of FA for both the A.1 and A.2 data matrices are presented as follows. Table 6 shows RE values and the study of the significance of the eigenvalues through an F-test. For the A.1 data matrix, Fig. 1 plots the evolution of IND function vs. number of factors and Fig. 2 the evolution of SEP (standard error or prediction) functions vs. number of factors, evaluated according to Malinowski and Wold cross-validation techniques.

For both the A.1 and A.2 data matrices, all the methods used lead to the confirmation that two factors are sufficient to explain the variation of the studied data sets. Hence this number of factors was adopted for the target testing.

Table 7 shows the results of testing the Kamlet and Taft parameters, with n_2 and unity as targets. The target vector unity must be included in the target set in order to take into account the chemical nature of each substance, this nature being a constant contribution in all data referring to each substance. For the A.1 data matrix n_2, π^* and unity were accepted as targets lying in

TABLE 6

Determination of the number of factors, NF

NF	A.1 data matrix				A.2 data matrix			
	RE [a]	Log(EV)	F(calc.)	SL(%) [b]	RE [a]	Log(EV)	F(calc.)	SL(%) [b]
1	0.360	3.335	850.89	0.0	0.579	2.885	148.35	0.7
2	0.043	0.584	101.53	*0.01* [c]	*0.016* [c]	0.605	1067.06	*1.9* [c]
3	0.029	−1.519	1.75	27.8		−2.821		
4	0.019	−1.979	1.25	37.9				
5	0.008	−2.378	2.22	37.6				
6		−3.325						

[a] RE = real error, according to [22]. [b] SL = percentage significance level associated with each F-value, calculated using log(eigenvalues) [log(EV)] according to [24]. [c] Values in italics indicate the correct number of factors according to the different criteria.

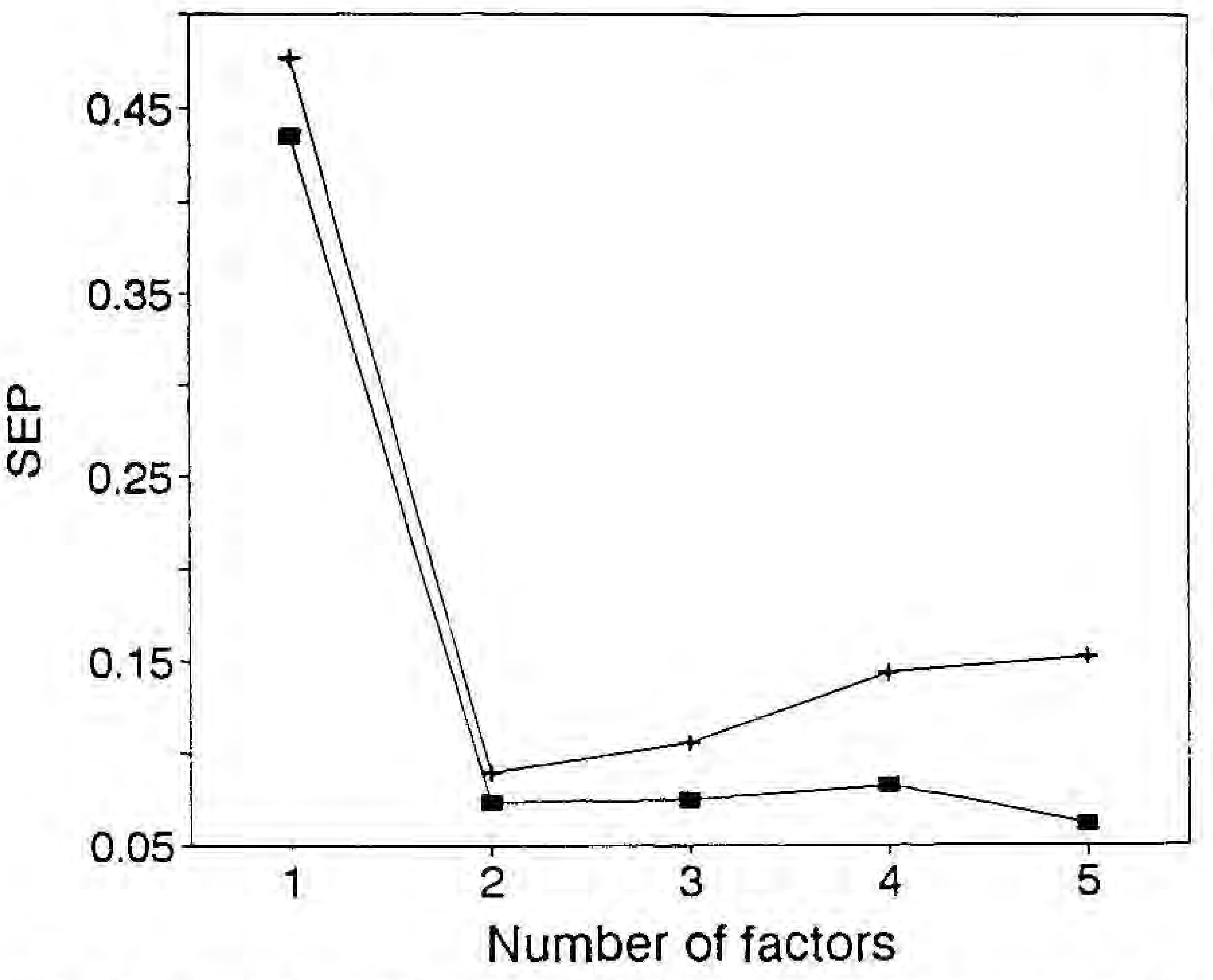


Fig. 2. Evolution of the SEP functions according to (■) Malinowski and (+) Wold vs. the number of factors for the A.1 data matrix.

the factor space. Of the two reasonable combinations of targets, π^*–unity and n_2–unity, the former reproduces the data matrix with an RMS = 0.052, whereas the latter yields an RMS = 0.060. Even though the difference does not seem very significant, the results of target testing for the A.2 data matrix confirm clearly the superiority of the π^*–unity over the n_2–unity combination. When the composition range includes a higher percentage of dioxane, n_2 is not admitted as a correct target, and only π^* and unity are considered to be valid. For the reproduction of the A.2 data matrix, the combination of π^* and unity yields an RMS = 0.060. Definite loadings [which are defined as coefficients related to the weight of each real factor in each different column (each different substance) of the original data matrix] for the reproduction of the A.1 and A.2 data matrices are given in Table 8.

TABLE 7

Summary of target errors using two factors based on covariance for A.1 and A.2 data matrices

Target	A.1 data matrix [a]						
	AET	EDM	SPOIL	F	df1	df2	SL (%)
α	3.59×10^{-2}	6.55×10^{-3}	5.39	9.14	4	4	2.7
β	4.55×10^{-2}	2.19×10^{-3}	20.78	131.50	4	4	0.0
π^*	1.76×10^{-2}	8.15×10^{-3}	1.92	1.42	4	4	37.1
Unity	8.91×10^{-3}	3.50×10^{-3}	2.30	1.97	4	4	26.4
n_2	1.33×10^{-2}	4.72×10^{-3}	2.63	2.40	4	4	20.8
	A.2 data matrix [a]						
	AET	EDM	SPOIL	F	df1	df2	SL (%)
α	3.51×10^{-2}	2.91×10^{-3}	11.98	55.30	4	1	10.0
β	4.73×10^{-2}	9.70×10^{-4}	48.80	911.59	4	1	2.5
π^*	9.08×10^{-3}	3.77×10^{-3}	2.19	2.22	4	1	46.1
Unity	5.07×10^{-3}	1.83×10^{-3}	2.58	2.93	4	1	40.9
n_2	3.27×10^{-2}	1.59×10^{-3}	20.56	162.16	4	1	5.9

[a] AET = apparent error in the target; EDM = error from data matrix; SPOIL = spoil associated with the target; df1 an df2 = degrees of freedom; SL = percentage significance level.

TABLE 8

Factor loadings for the reproduction of the A.1 and A.2 data matrices
(Number of factors = 2. Factor loadings based on covariance)

Compound [a]	A.1 data matrix [b]		A.2 data matrix [b]	
	π^*	Unity	π^*	Unity
PropK	−6.4(2)	12.4(2)		
Sal1	−5.8(2)	9.5(2)		
Sal2	−5.4(4)	19.3(4)		
HNCA	−6.3(1)	9.9(1)	−6.7(2)	10.5(2)
Gly1	−4.38(9)	7.5(1)	−4.9(2)	8.1(2)
Gly2	−1.18(9)	10.95(9)	−1.6(2)	11.4(1)

[a] For abbreviations, see Table 3. [b] Values in parentheses are the errors associated with the last figures.

This study confirms the greater usefulness of microscopic parameters, such as π^*, over bulk parameters, such as n_2, in the explanation of microscopic processes, as the solvent properties in the cybotactic zone, which are often very different from those in the bulk solvent, are the ones which affect directly the solutes when a process such as acid–base equilibrium occurs.

The application of FA and TFA to the study of the effect of solvent properties on the solute protonation constants in water–dioxane mixtures confirms the model suggested in the classical attempt for most substances, leading to the reduced form of the Kamlet and Taft equation $pK = pK_0 + s\pi^*$, independent of the functional group and composition range studied.

Only for propionic acid, determined at $I = 0.2$ M in KNO_3, should the α parameter also be introduced into the model, according to the step-

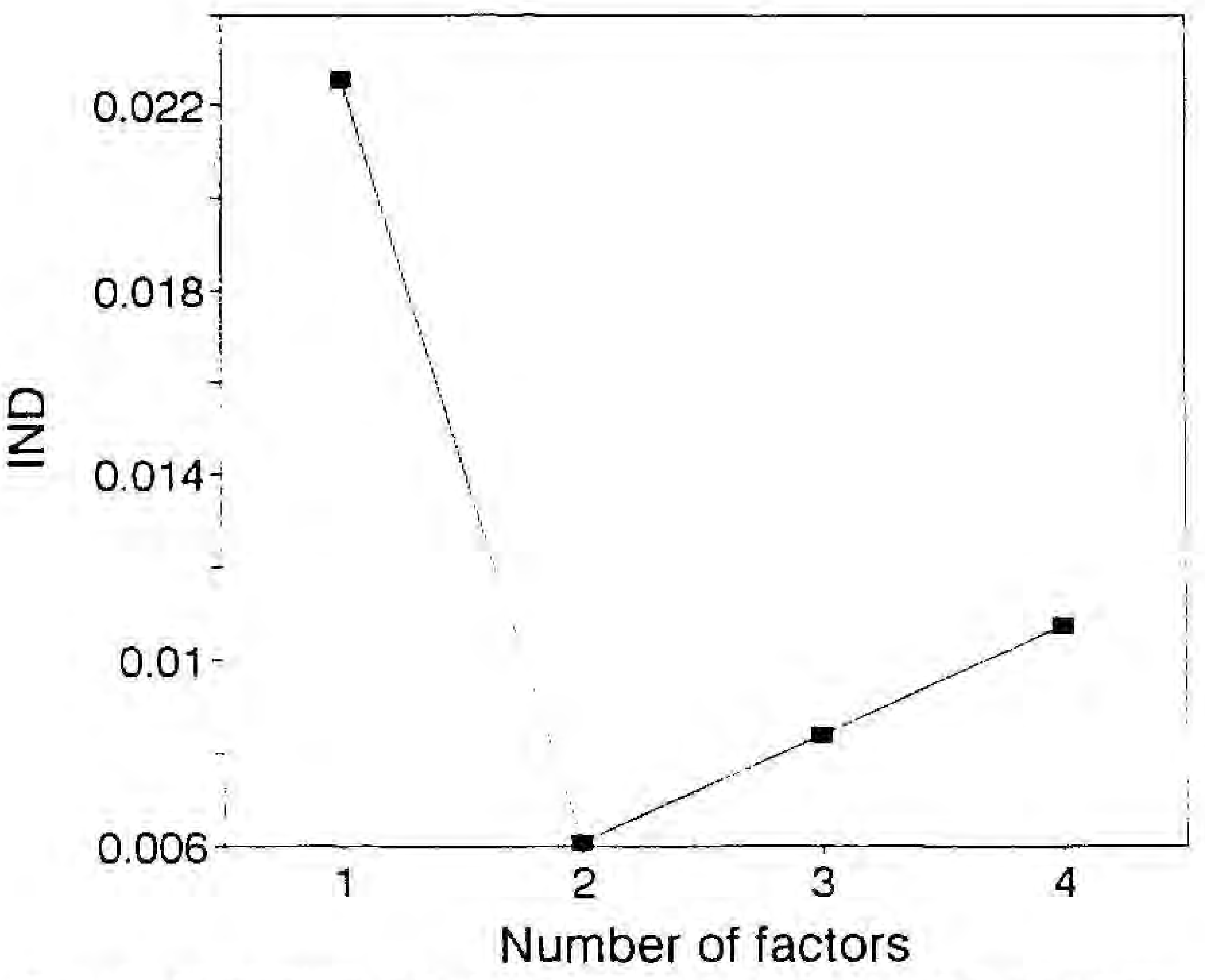


Fig. 3. Evolution of the IND function vs. the number of factors for the B.2 data matrix.

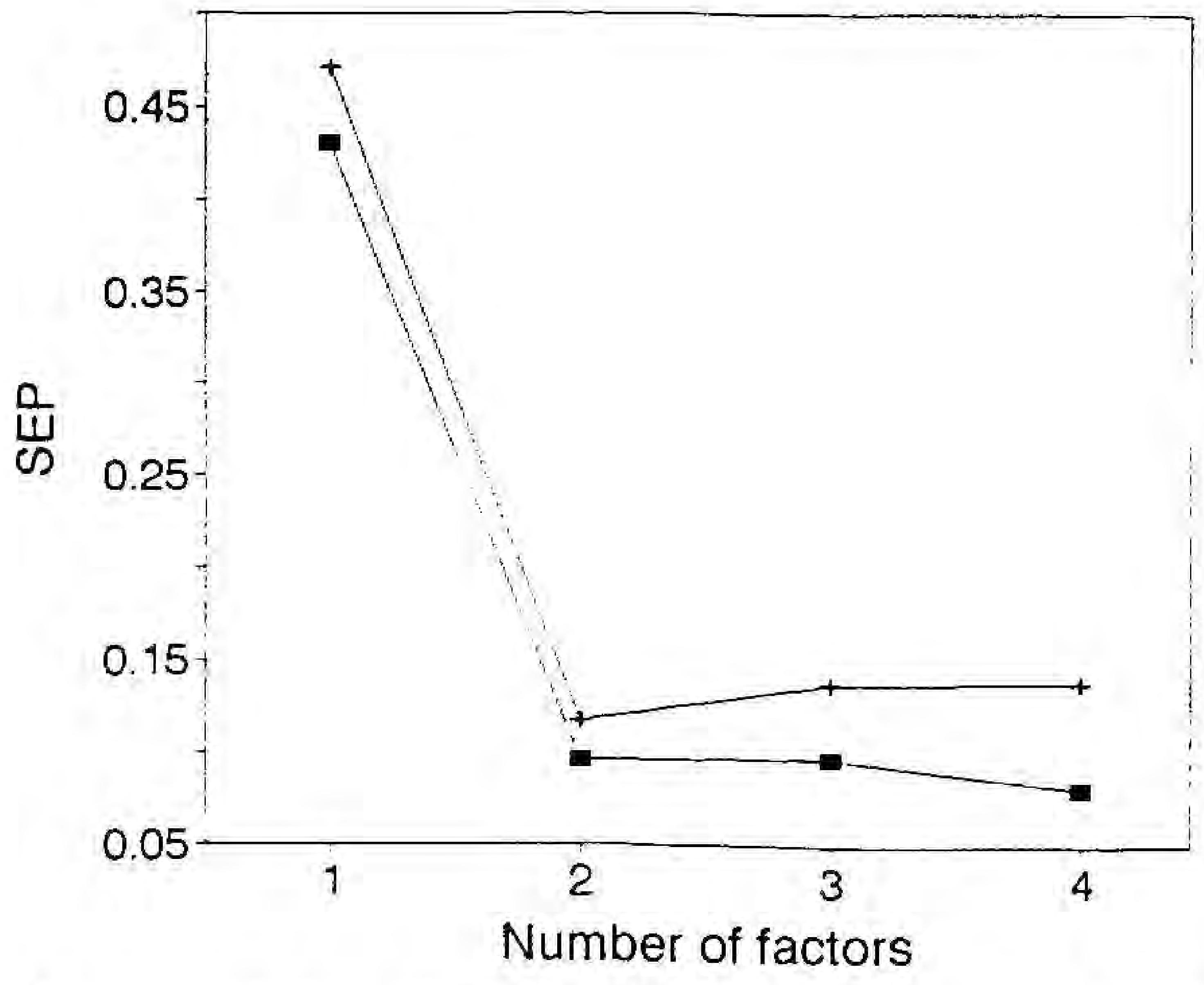


Fig. 4. Evolution of the SEP functions according to (■) Malinowski and (+) Wold vs. the number of factors for the B.2 data matrix.

wise procedure. This difference in the model is actually of minor significance, because π^* explains 98.9% of the total variance in the data, and the introduction of α increases the level of explanation to 99.8%. In fact, the expression obtained including only π^* yields a high-quality fit: $pK_s = (-6.5 \pm 0.2)\pi^* + (12.5 \pm 0.2)$; $r^2 = 0.996$; scale estimate = 0.05.

The subtle numerical differences between loadings and correlation coefficients for every substance are easily explained, as the amount of data used for establishing the two parallel models is not always the same and in the case of the amino group of glycine one outlier has been removed. Of the expressions obtained by the two methods, individual models are to be preferred, because they are built using a larger number of quality-checked data.

The results obtained in the study of the effect of the inert electrolyte on the protonation constants does not lead to very clear inferences. There is no evident variation in the number of factors when data determined with a different electrolyte are added to the B.1 data matrix, as can be seen from Table 9 and Figs. 3 and 4. As the size of the B.1 data matrix used is small (other work is in progress), the only tentative conclusion that can be extracted from the present data sets is that no apparent electrolyte effect can be observed or, at least, that if it is present, its magnitude is so small that it is masked by some other main effects.

TABLE 9

Determination of the number of factors, NF

NF	B.1 data matrix				B.2 data matrix			
	RE [a]	Log(EV)	F(calc.)	SL(%) [b]	RE [a]	Log(EV)	F(calc.)	SL(%) [b]
1	0.364	3.011	536.19	0.0	0.362	3.263	777.54	0.0
2	*0.050* [c]	0.197	42.79	*2.3* [c]	*0.055* [c]	0.489	57.45	*0.5* [c]
3	0.007	−1.695	22.57	13.2	0.034	−1.396	1.97	29.5
4		−3.650			0.011	−1.891	6.11	24.5
5						−3.153		

[a] RE = real error, according to [22]. [b] SL = percentage significance level associated with each F-value, calculated using log(eigenvalues) [log(EV)] according to [24]. [c] Values in italics indicate the correct number of factors according to the different criteria.

This research was supported by the Spanish Comision Interministerial de Ciencia y Tecnologia (CICYT), Grant No. PB90-0821, and by a grant from the Generalitat de Catalunya (CIRIT).

REFERENCES

1 E.R. Malinowski and D.G. Howery, Factor Analysis in Chemistry, Wiley, New York, 1980.

2 P.H. Weiner, J. Am. Chem. Soc., 95 (1973) 5845.

3 W.R. Fawcett and T.M. Krygowski, Can. J. Chem., 54 (1976) 3283.

4 J.T. Edward and S.C. Wong, J. Am. Chem. Soc., 99 (1977) 4229.

5 E.M. Kosower, An Introduction to Physical Organic Chemistry, Wiley, New York, 1968.

6 E. Casassas, G. Fonrodona, A. de Juan and R. Tauler, Chemometr. Intell. Lab. Syst., 12 (1991) 29.

7 M.J. Kamlet, J.J. Abboud and R.W. Taft, J. Am. Chem. Soc., 99 (1977) 6027.

8 R.W. Taft and M.J. Kamlet, J. Am. Chem. Soc., 98 (1976) 2886.

9 M.J. Kamlet and R.W. Taft, J. Am. Chem. Soc., 98 (1976) 377.

10 A.I. Vogel, A Text Book of Practical Organic Chemistry, Longmans Green, London, 5th edn., 1989, p. 407.

11 J.E. Powell and M.A. Hiller, J. Chem. Educ., 34 (1957) 330.

12 E. Casassas and G. Fonrodona, J. Chim. Phys., 86 (1989) 391.

13 E. Casassas, G. Fonrodona and A. de Juan, Inorg. Chim. Acta, 187 (1991) 187.

14 G. Gran, Acta Chem. Scand., 4 (1950) 559; Analyst, 77 (1952) 661.

15 P. Gans, A. Sabatini and A. Vacca, J. Chem. Soc., Dalton Trans., (1985) 1195.

16 K.K. Mui, W.A.E. McBryde and E. Nieboer, Can. J. Chem., 52 (1974) 1821.

17 E. Casassas, G. Fonrodona and A. de Juan, J. Solution Chem., 21 (1992) 147.

18 C. Reichardt, in H. Ratajczak and W.J. Orville-Thomas (Eds.), Molecular Interactions, Vol. 3, Wiley, Chichester, 1982, p. 241.

19 M.J. Kamlet and R.W. Taft, Acta Chem. Scand., Ser. B, 39 (1985) 611.

20 M. Forina, R. Leardi, C. Armanino and S. Lanteri, PARVUS, an Extendable Package of Programs for Data Exploration, Classification and Correlation, Elsevier, Amsterdam, 1988.

21 P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1981.

22 E.R. Malinowski, J. Chemom., 1 (1987) 33.

23 S. Wold, Technometrics, 20 (1978) 397.

24 E.R. Malinowski, J. Chemom., 3 (1988) 49.

25 E.R. Malinowski, Anal. Chim. Acta, 103 (1978) 339.