Journal of Chromatography, 599 (1992) 131–136
Elsevier Science Publishers B.V., Amsterdam

CHROMSYMP. 2555

Characterization of hydrophobic interaction and hydrophobic interaction chromatography media by multivariate analysis

Per Kårsnäs* and Tua Lindblom

Pharmacia BioProcess Technology AB, S-751 82 Uppsala (Sweden)

ABSTRACT

Mapping of the selectivity of different hydrophobic interaction chromatography (HIC) media was performed using principal component analysis. The elution positions of proteins in a descending salt gradient were determined for a selection of commercially available and experimental gels. Data regarding the proteins obtained from the literature, such as hydrophobicity indices, polar/non-polar volume ratio and charge, were compared with the elution positions. Principal component analysis revealed that the media could be divided into several groups. The map also showed that the different hydrophobicity indices could be related to the different retention patterns of the media groups. Some media were selecting mainly according to protein hydrophobicity as expressed by the fraction of hydrophobic amino acids, and their mechanism of interaction was insensitive to protein charge. On other media the charge on the protein, or rather the lack of charge, was the most important factor. The results suggest that HIC in many cases works by a combination of two different mechanisms, thus explaining the irregular behaviour which is so often experienced when using HIC media. Principal component analysis and other multivariate techniques were found to be valuable tools in the process of understanding and optimizing hydrophobic or salt-promoted interaction chromatography.

INTRODUCTION

Hydrophobic interaction chromatography (HIC) is a separation technique that is widely used in bioscience and biotechnology. Sometimes, users find that the behaviour of proteins on HIC media is unpredictable. Generally, it is difficult to identify the characteristics of a protein, *e.g.* its hydrophobicity index, that will predict its interaction with a HIC matrix, in the way that the isoelectric point is used in the case of ion-exchange chromatography.

Our study began with the aim of empirically characterizing hydrophobic matrices by determining the elution positions of a set of test proteins in two different standard gradients. In this way we anticipated obtaining a map of the differences in the elution selectivity of the tested media. Later we found it a natural approach to process the generated information by multivariate analysis, a tool that we had started to investigate and use in other contexts. This then permitted us to directly compare other protein data, such as hydrophobicity indices (HIs), hydrophilicity coefficients, polar/non-polar volume ratio and charge, with the chromatographic data.

EXPERIMENTAL

Materials

The experiments were performed using fast protein liquid chromatography (Pharmacia LKB Biotechnology, Uppsala, Sweden). A bed height of 10 cm was packed in HR 10/10 columns (Pharmacia). The proteins were divided into two separate sample mixtures: 1, myoglobin, β-lactoglobulin and α-lactalbumin; 2, cytochrome c, ribonuclease, lysozyme and α-chymotrypsinogen A (all chemicals from Sigma). The protein concentrations were 1 mg/ml (β-lactoglobulin 2 mg/ml) and the sample volume 1 ml. Two gradients with different starting conditions

0021-9673/92/$05.00

TABLE I

INVESTIGATED HIC MEDIA

Agarose–hexane (AgHex)	Phenyl Sepharose 6 FF (low sub) (PheFFLs)
Butyl Sepharose 4B (But4B)	Phenyl Sepharose 6 FF (high sub) (PheFFhs)
Butyl Sepharose 4 FF (ButFF)[a]	Neopentyl HP (Nphp)[b]
Octyl Sepharose CL-4B (OctC14B)	Thio ether Sepharose 6 FF (TEFF)[b]
Octyl Sepharose 6 FF (OctFF)[a]	Hexyl mercaptane 6 FF (HexMFF)[b]
Phenyl Sepharose CL-4B (Phe4B)	Pyridine sulphate 6 FF (PySFF)[c]
Phenyl Sepharose HP (PheHP)	

[a] Only available as custom designed medium from Pharmacia.
[b] Experimental gels. Not commercially available.
[c] To be launched.

were used for every protein. The proteins were dissolved in the starting buffer, 100 m*M* sodium phosphate pH 7.0 containing 1 *M* or 1.5 *M* ammonium sulphate. The final buffer was 100 m*M* sodium phosphate, pH 7.0. The gradient volume was ten column volumes, the linear flow-rate 40 cm/h for Sepharose 4B and Sepharose CL-4B matrices, and 75 cm/h for Sepharose FF matrices.

Media

The tested media are listed in Table I. Seven of them are commercially available, two at present only in bulk as custom-designed media from Pharmacia. One medium, pyridine sulphate Sepharose FF, is to be launched, and three media are not commercially available and were tested to increase the base of the characterization (Table I).

Collection of data

Two elution positions of the seven proteins, one for each gradient, were determined for every chromatographic medium. The void volume was subtracted. The values of five different. hydrophobicity (or hydrophilicity) indices, the polar/non-polar volume ratio (p) [1] and charge [1] completed the variables. The HIs were: the non-polar side chain frequency (NPS) [1], hydrophilicity according to Hopp and Woods [2], and hydrophobicity according to Janin [3], Kyte and Dolittle [4] and Rose *et al.* [5]. The data are summarized in Table II.

Multivariate an analysis

General description [6]. Any data table built up of x rows and y columns can be represented by y vectors in an x-dimensional space. Such a vector graphically carries exactly the same information as the table. If, for example, $x = 3$ and $y = 3$, the resulting three-dimensional diagram is easy to interpret. On the other hand, it is very difficult to imagine the nineteen vectors (variables) in the fourteen-dimensional space (experiments, objects) which is the result of this investigation. Mathematically, however, the passage from three to four dimensions and more is trivial.

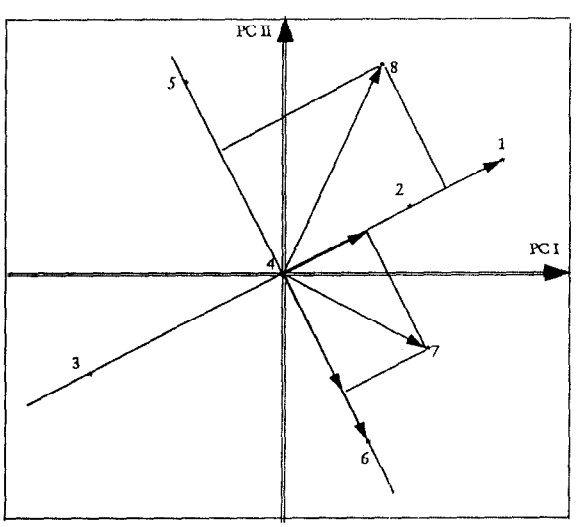

Fig. 1. Interpretation of a principal component projection of variables. Consider point 1 to correspond to a dependent variable. Variable 2 is covariate and 3 varies inversely to 1. Points 4, 5 and 6 vary independently of 1. If 6 is considered a second dependent variable, an increase in variable 7 will increase the values of both 1 and 6, while an increase in 8 will increase 1 and decrease 6. For further details see the text.

In multivariate analysis interpretation of the information in the data set is made possible by projections of the multidimensional space on to planes. The planes which retain most of the available information can be determined by principal component analysis (PCA). The first principal component (PC I) is the direction in the multidimensional space in which the sum of the lengths of the projections of all the vectors is the highest. The length of the projection of a variable vector on a PC is called "loading". The second PC is the direction in which the sum of projections is the next highest, and moreover it has to be perpendicular to the first PC. In the same way more PCs can be calculated. The later principal components carry less and less information, and in practice it is often found that from the fourth PC onwards only noise is expressed. If the number of PCs equals the number of dimensions in the experimental space, the only result of the operation is a transformation of the axes in the space. Very often, in practice, the plane spanned by the first and second PGs carries more than 75% of the

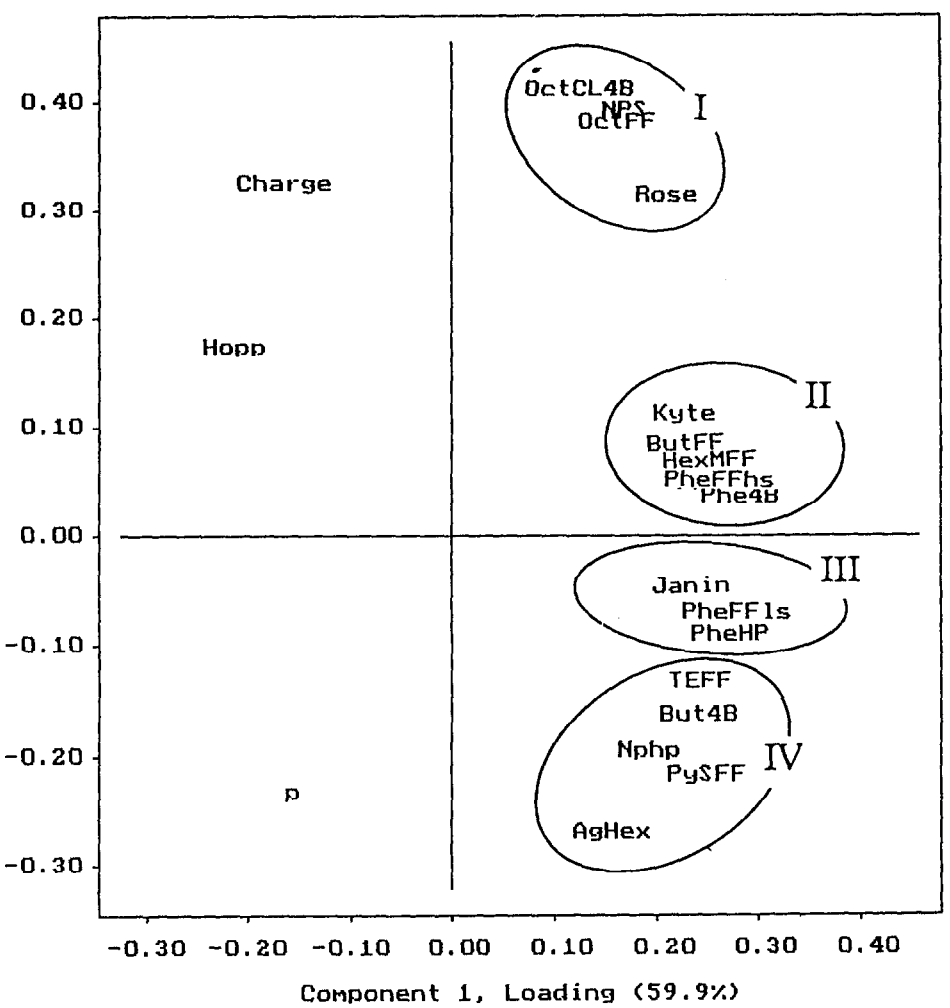

Fig. 2. Principal component map of the selectivity of HIC media (loadings). A classification of the media into different groups is indicated. Some media (I) retard proteins mainly depending on their content of hydrophobic amino acids, some (IV) in inverse proportion to charge and some (II, III) by a combination of the two. The existence of at least two different interaction mechanisms is visualized. The covaration of the media groups with the different indices is clearly vizualized.

TABLE II

EXPERIMENTAL DATA

Objects: proteins at two different starting conditions. Variables: retention times on the different media, hydrophobicity indices (Rose, Kyte and Janin), hydrophilicity coefficients (p and Hoppe) and charge. Proteins that stick to the column at the final gradient conditions were given the retention time 100. A variation of this figure from 80 to 200 did not significantly change the results.

Protein[a]	Medium																		
	AgHex	ButFF	OctCL4B	OctFF	PheHP	Phe4B	PheFFls	PheFFhs	Nphp	TEFF	HexMFF	PySFF	NPS	p	charge	Rose	Kyte	Janin	Hopp
Cyt c	1	0	1	1	0	1	0	0	0	0	0	0	27	112	34	0.54	−0.92	−0.33	0.38
Cyt c	1	1	1	2	1	2	1	2	0	0	1	1	27	112	34	0.54	−0.92	−0.33	0.38
Myoglobin	1	1	100	100	2	12	2	14	0	0	1	2	32	112	34	0.64	−0.40	−0.10	0.14
Myoglobin	2	2	100	100	24	26	11	11	1	1	9	6	32	112	34	0.64	−0.40	−0.10	0.14
RNAse	1	18	5	1	6	7	5	31	0	0	7	5	23	173	24	0.48	−0.67	−0.12	0.09
RNAse	14	19	12	12	26	18	13	26	2	0	8	17	23	173	24	0.48	−0.67	−0.12	0.09
Lysozyme	1	18	19	28	27	39	25	53	0	6	9	31	26	118	14	0.64	−0.48	0.00	−0.04
Lysozyme	14	41	33	49	49	52	41	65	6	17	31	45	26	118	14	0.64	−0.48	0.00	−0.04
β-Lac	1	45	100	100	16	47	21	61	0	0	26	2	37	96	28	0.70	−0.16	−0.06	0.11
β-Lac	6	62	100	100	41	57	38	66	1	8	53	20	37	96	28	0.70	−0.16	−0.06	0.11
α-Lac	1	78	100	100	49	47	31	81	0	6	26	28	34	111	28	0.71	−0.46	−0.06	0.10
α-Lac	13	81	100	100	64	61	45	81	9	22	53	43	34	111	28	0.71	−0.46	−0.06	0.10
α-Chy	1	46	26	43	55	56	39	71	0	6	18	48	33	83	15	0.64	0.05	0.12	−0.26
α-Chy	14	60	40	59	68	65	53	77	18	17	42	60	33	83	15	0.64	0.05	0.12	−0.26

[a] Cyt c = Cytochrome c; β-Lac = β-lactoglobulin; α-Chy = α-chymotrypsinogen A.

information of the entire multidimensional space.

It is also possible to let the variable vectors form the space and the object vectors be expressed in this space. In this case the projections are called scores. A PC plane in this space will express the quality of the experimental design, *e.g.* whether or not the space is properly spanned by the values of experimental parameters.

Interpretation of PC projections. In Fig. 1 is shown an example of a PC projection plane. If point 1 corresponds to a dependent variable, the direction from the point through the origin is of great interest. The independent variable 2 is pointing in the same direction, which means that increasing the value of the corresponding parameter will also increase the value of parameter 1. Variable 3 points in the opposite direction and an increase in the value of this parameter will decrease the value of 1! Parameter 4, which is situated near the origin, does not have any influence at all. It can be varied freely or, for example, if the set-up is a quality control, it does not have to be measured at all. If the scales of the two PC axes are the same, any parameter on a line through the origin perpendicular to the direction of 1 will have no influence on 1.

Analysis of HIC data. A protein combined with

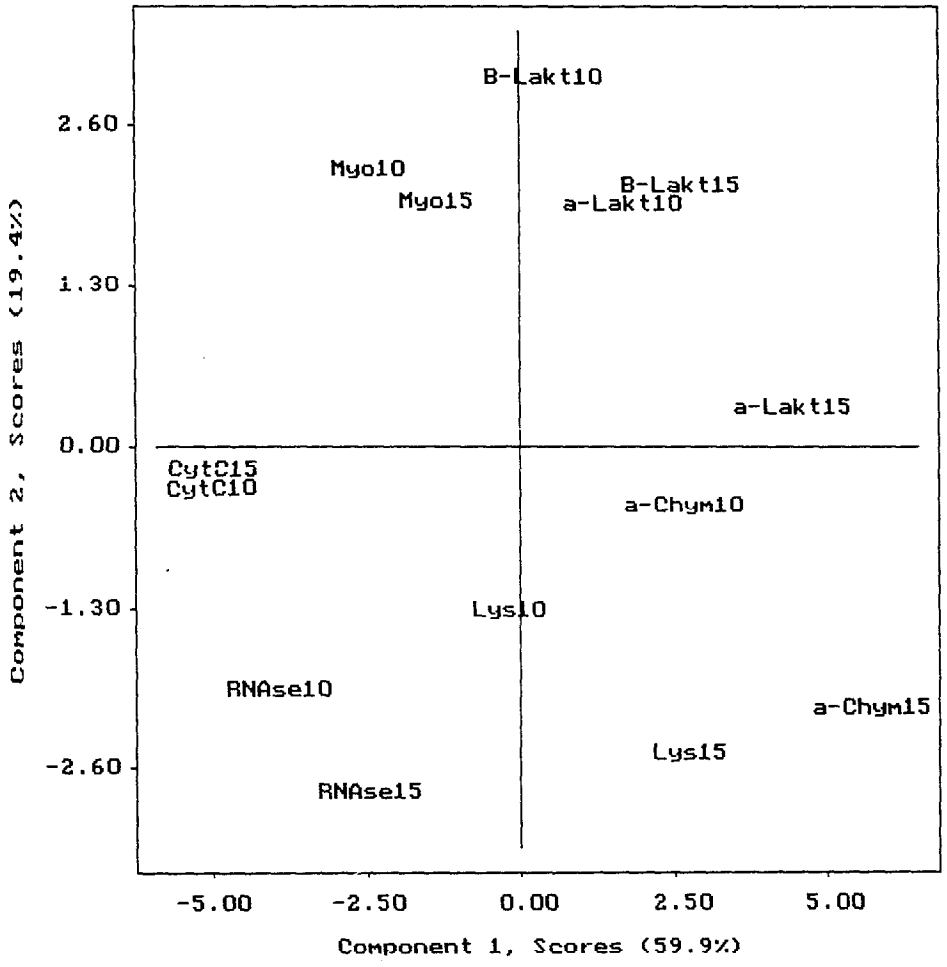

Fig. 3. Principal component projection (score projection) of the proteins with the initial concentration of ammonium sulphate at which they were run (1 *M* ammonium chloride designated "10", etc.). The diagram shows that the objects span the experimental space evenly, thus expressing high reliability of the variable projection.

one starting concentration of the gradient was considered an object (experiment). Thus, every protein gave rise to two objects. The values of the corresponding elution positions and also the values of the hydrophobicity (hydrophilicity) indices (Rose, Kyte, Janin, Hoppe), the polar/non-polar ratio (p) and the charge on the test proteins were variables. The resulting table was fed into SIRIUS (Pattern Recognition Systems, Bergen, Norway), a computer program for multivariate calculations. The data set was centered and the variable values standardized by dividing by the standard deviation. Three PCs were calculated.

RESULTS AND DISCUSSION

The loading plane formed by PC I and PC II was found to carry 80% of the information in the data set. The resulting diagram (Fig. 2) reveals that the media can be divided into several groups or classes depending on their elution selectivity pattern. Moreover the different HIs were shown to be covariate with different classes of gels. The variation in the retention on the octyl gels of group I is covariate with the NPS values [1] of the proteins and also with the HI as defined by Rose *et al.* [5]. Since this variation is nearly perpendicular to the charge, the difference in the charge on the proteins is of less importance for the binding to octyl gels. The media in group IV retard the proteins more the less their charge, because the corresponding vectors are in the opposite direction. The NPS value is of less importance for this interaction. The two groups mentioned clearly point towards the existence of two separate mechanisms of HIC, one related to the fraction of hydrophobic amino acids of a protein and one inversely related to protein charge. One way of interpreting this is that the first type of interaction occurs some distance inside the molecule while the second interaction only occurs with hydrophobic structures at the surface where the interaction will be increased by the absence of charges in the vicinity.

Thus most media of phenyl and butyl type work according to a mixed mode where both hydrophobicity and lack of charge simultaneously play a role. Of course, this may explain the sometimes unpredictible behaviour of HIC media. It is interesting that the HIs defined by Kyte and Dolittle [4] and Janin [3] correlate well with the retention on group II and group III media, respectively. Obviously they express hydrophobicity in a more functional way than NPS, which is just describing the fraction of hydrophobic amino acids without taking any account of whether or not they are exposed on the surface of the protein. Further experiments with other proteins have to be performed to verify which of the indices is best able to predict how proteins will interact with the different types of media. The score projection (Fig. 3) shows that the chosen proteins span the experimental space in a proper way.

There are other possible ways of using a PCA map of the kind presented here. By running the test proteins on other HIC media in the same way it is possible to classify these and to find possible substitutes for a certain medium. Very often we have found that, for example, the butyl media made by one manufacturere have selectivities more similar to the phenyl media of another. The reason for this is probably differences in the type of base matrix, coupling chemistry and spacers.

If a certain protein is run on a sufficient number ($\geqslant 5$) of the media in the map it is possible to classify it as similar to one of the proteins tested here. Successful separation of the test protein will then also be valid for the unknown protein. To establish this as a successful optimization method, further work is required.

ACKNOWLEDGEMENT

The authors are indepted to Dr. Rolf Hjort, Pharmacia BioProcess Technology AB, for help with the calculation of hydrophobicity coefficients.

REFERENCES

1 C. C. Biglow, *J. Theor. Biol.*, 16 (1967) 187.
2 T. P. Hopp and K. R. Woods, *Proc. Natl. Acad. Sci. USA*, 78 (1981) 3824.
3 J. Janin, *Nature*, 277 (1979) 491.
4 J. Kyte and R. F. Dolittle, *J. Mol. Biol.*, 157 (1982) 105.
5 G. D. Rose, L. M. Gierasch and J. A. Smith, *Adv. Protein Chem.*, 37 (1985) 1.
6 S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg and H. Sjöström, in B. R. Kowalski (Editor), *Chemometrics: Mathematics and Statistics in Chemistry*, Reidel, Dordrecht, 1984, Ch. 1, p. 1.