Large contact surface interaction between proteins detected by time series analysis methods: case study on C-Phycocyanins

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MAY 2003

Impact Factor: 2.63 · DOI: 10.1002/prot.10366 · Source: PubMed

5 AUTHORS, INCLUDING:

Alessandro Giuliani
Istituto Superiore di Sanità
362 PUBLICATIONS **4,461** CITATIONS
SEE PROFILE

Indu R Chandrashekaran
Monash University (Australia)
19 PUBLICATIONS **216** CITATIONS
SEE PROFILE

Sudha Cowsik
32 PUBLICATIONS **413** CITATIONS
SEE PROFILE

Large Contact Surface Interactions Between Proteins Detected by Time Series Analysis Methods: Case Study on C-Phycocyanins

Alessandro Giuliani,[1] Romualdo Benigni,[1,*] Mauro Colafranceschi,[1,2] Indu Chandrashekar,[3] and Sudha M. Cowsik[3]

[1]*Istituto Superiore di Sanita', Lab. TCE, Rome, Italy*
[2]*University of Rome "La Sapienza," Physiology and Pharmacology Department, Rome, Italy*
[3]*Jawaharlal Nehru University, School of Life Sciences, New Delhi, India*

ABSTRACT **A purely sequence-dependent approach to the modeling of protein–protein interaction was applied to the study of C-phycocyanin αβ dimers. The interacting pairs (α and β subunits) share an almost complete structural homology, together with a general lack of sequence superposition; thus, they constitute a particularly relevant example for protein–protein interaction prediction. The present analysis is based on a description posited at an intermediate level between sequence and structure, that is, the hydrophobicity patterning along the chains. Based on the description of the sequence hydrophobicity patterns through a battery of nonlinear tools (recurrence quantification analysis and other sequence complexity descriptors), we were able to generate an explicit equation modeling α and β monomers interaction; the model consisted of canonical correlation between the hydrophobicity autocorrelation structures of the interacting pairs. The general implications of this holistic approach to the modeling of protein–protein interactions, which considers the protein primary structures as a whole, are discussed. Proteins 2003; 51:299–310.** © 2003 Wiley-Liss, Inc.

Key words: recurrence quantification analysis; singular value decomposition; bioinformatics; principal component analysis; protein–protein interaction

INTRODUCTION

Phycobilisomes , the light-harvesting complexes of cyanobacteria and red algae, are large multimeric protein structures located on the thylakoid membrane. They are optimized for light absorption and transfer of energy via the phycobilisome core to the membrane-bound photosynthetic reaction center.[1,2] Each phycobilisome contains a core and a number of rods, made of stacks of hexamers of dimers of phycocyanin molecules; the dimers in turn consist of two structurally homologous proteins α and β[2,3] (Fig. 1). The assembly of phycobilisomes typically goes through the subsequent steps of trimerization of dimers $(\alpha\beta)_3$ and formation of hexamers $(\alpha\beta)_6$, the basic unit of the complex being the αβ dimer. The phycobilisomes contain varying ratios of different members of this protein family: phycoerythrin, phycocyanin, allophycocyanin, and a number of other minor variants. The particular composition of the phycobilisomes is linked to the peculiar light absorption spectrum. The rods also contain linker proteins situated in the central trimer cavity. However, self-assembly of phycocyanins into hexamers occurs in vitro in the absence of linker proteins as well, with the two α and β units interacting in a specific way and along a large contact surface.[1–3] This fact makes this protein family an ideal model for studying protein–protein interaction by holistic methods, taking into consideration the entire protein sequence and not only specific zones of the primary structure.

Moreover, given the presence of strict structural and dynamic constraints shaping the phycobilisome assembly, the α and β subunits interaction has to be highly specific. Thus, the identification of a proper level of analysis of the correlation between α and β subunits of different species assumes a far-reaching value for the elucidation of protein–protein interaction process. The pairs of α and β elements of the C-phycocyanins, while having practically the same 3D structure (Fig. 1), do not share any significant sequence homology (Fig. 2). Thus, neither 3D structure nor sequence homology provide an immediate reason for the interaction. In other words, while the identity of the 3D structures of the α and β monomers does not offer us a simple reason for the necessity of having a heterodimer instead of an homodimer to build a functioning phycobilisome, the absence of any relevant sequence homology between the two subunits does not offer any cue to model, on a pure sequence alignment basis, the observed interaction. This implies the correct level of analysis of the observed interaction should be situated at some intermediate level between sequence and structure .

A large body of evidence points to the reciprocal similarity of hydrophobicity distribution patterns as the most important driving force for protein–protein interaction, as

Grant sponsor: NSF/NIH; Grant number: 0240 230

*Correspondence to: Romualdo Benigni, Istituto Superiore di Sanita', Lab. TCE, Viale Regina Elena 299, 00161 Rome, Italy. E-mail: rbenigni@iss.it

Received 28 October 2002; Accepted 22 November 2002

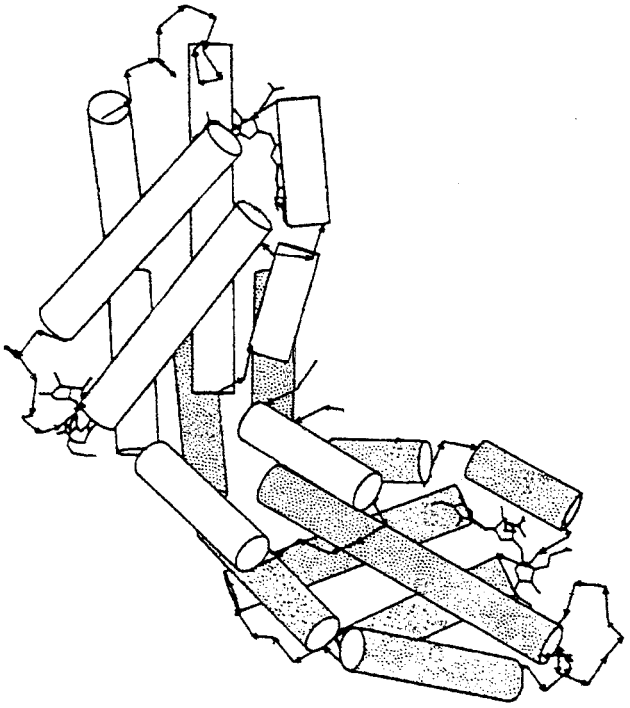

Fig. 1. 3D structure of *Synechococcus agmenellum* quadruplicatum (agme) c-phycocyanin. The cylinders are α helices. Grey cylinders: α-chain. White cylinders: β-chain.

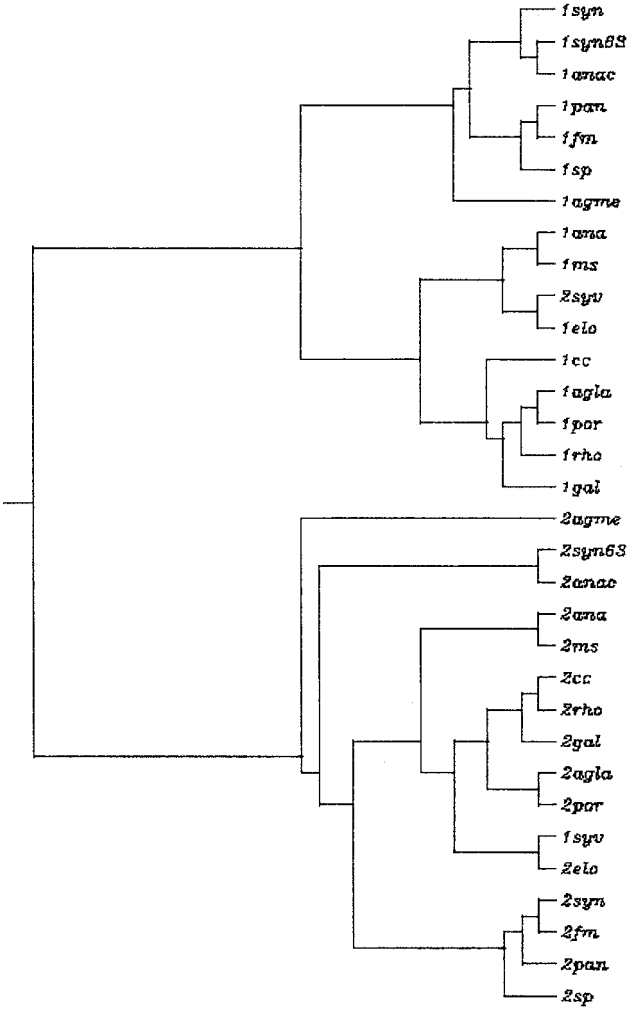

Fig. 2. ClustalW multiple alignement tree of c-phycocyanins (α and β chains) is reported. 1, α; 2, β. Note that the relative positions in the tree of α and β chains of *syv* are inverted.

well as for protein folding.[4–6] From a modeling point of view, the sequential arrangement of the amino acidic side-chain hydrophobicity can be paralleled to a time series, where the residue positions play the role of time. Thus, the descriptions offered by the time series analysis tools for quantifying the autocorrelation properties of the series can be utilized for generating a quantitative description of the hydrophobicity patterns along the chain. Such a strategy has already demonstrated its heuristic power in a wide range of studies (for a review see [7]). In the case of protein–protein interaction, this approach has demonstrated its validity in several studies, from peptide–receptor to chaperone–protein interaction modeling.[4,5,8,9] These successful approaches used different signal analysis methods, like singular value decomposition (SVD) and recurrence quantification analysis (RQA). In this work we investigated the hypothesis that a battery of nonlinear descriptors of time series autocorrelation structures—already demonstrated optimal for the estimation of the degree of complexity of diverse time series[10]—is able to point to pairwise correlations of the autocorrelation structures of interacting pairs of protein sequences. The demonstration of a correlation between the hydrophobicity patterns of the interacting pairs can be considered as the "counterpart" of the interaction process between the two α and β molecules. This turned out to be actually the case, so providing further support to the relevance of nonlinear signal analysis strategies for modeling protein sequences, and opening the way to speculative hypotheses on the protein oligomerization process. It is worth stressing that this correlation was obtained only on the basis of sequence information, without any explicit reference to structure, thus constituting a brand new approach to protein–protein interaction.

DATA AND METHODS

Data

Sixteen pairs of α and β C-phycocyanins for a total of 32 proteins, correspondent to all the complete sequences existing in the Swiss–Prot archives (http://www.expasy.ch/sprot) at the end of May 2002 were analyzed in this work (Table I).

Methods

The primary structures were coded in terms of the Schneider and Wrede hydrophobicity scales.[11]

To extract an exhaustive description of the hydrophobicity patterning along the chains, the resultant numerical series were submitted to the battery of the following nonlinear signal analysis tools. The need to have the

TABLE I. Organisms, Swiss–Prot codes, and Code Used in the Text

Organism	Swiss–Prot code (α β chains)	Code
Synechocystis	P20776; P20777	*syn*
Pseudoanabaena	Q52447; Q52446	*pan*
Anabaena	P07121; P07120	*ana*
Fremyella diplosiphon	P07122; P07119	*fm*
Cyanidium calidarium	O19910; O1909	*cc*
Mastigocladus laminosus	P00307; P00311	*ms*
Spirulina platensis	P72509; P72508	*sp*
Thermosynechococcus elongatus	P50032; P50033	*syv*
Synechococcus anacystis nidulans (P6301 strain)	P00308; P00312	*syn*63
Aglaothamnion neglectum	P28557; P28558	*agla*
Galdieria sulphuraria	P00306; P00311	*gal*
Porphyra purpurea	P51378; P51377	*por*
Rhodella violacea	Q36699; Q36698	*rho*
Synechococcus elongatus	P50032; P50033	*elo*
Synechococcus agmenellum quadruplicatum	P03943; P03944	*agme*
Synechococcus anacystis nidulans (R2 strain)	P13530; P06539	*anac*

hydrophobicity series described by a large number of variables comes from the fact no single descriptor provides a fully satisfactory picture of the autocorrelation features of the observed series. On the contrary, when the global information conveyed by different algorithmic views of the autocorrelation structure of the series is summarized into a synthetic score by the agency of multidimensional analysis, we obtain a more robust and reliable score for describing the analyzed series.

Embedding -based methods

At the basis of all these methods is the projection of the original monodimensional series into a multidimensional space constituted by subsequently lagged copies of the original sequence. This corresponds to the generation of the so-called embedding matrix (EM). The EM columns are, in this order, (a) the original series; (b) the series shifted of one amino acid; (c) the series shifted of two amino acids; (d) etc. until a dimension variable from four to eight consecutive shifts is reached. Thus, EM is a multivariate matrix whose rows (statistical units) are subsequent patches (or sliding windows) of amino acids with length equal to the embedding dimension and whose columns (statistical variables) are the whole sequence lagged by subsequent delays. EM is an $M \times N$ matrix, M being the number of amino acids minus the embedding dimension (the last amino acids are eliminated by the shifting of the series due to the embedding procedure) and N the embedding dimension.[12]

RQA. RQA is a relatively new nonlinear technique, originally developed by Eckmann et al.[13] as a purely graphical technique and then made quantitative by Webber and Zbilut.[14] The technique was successfully applied to a several different fields, including physiology,[14] molecular dynamics,[15] chemical reactions,[16] and, more recently, protein sequences.[7,8] Notwithstanding the recent development of RQA, the notion of recurrence at the basis of this technique is well established.[13]

For any ordered series (temporal or spatial), a recurrence is defined as a point that repeats itself. Because recurrences are simply tallies, they make no mathematical assumptions. Given a reference point, X_0, and a ball of radius r, in an N-dimensional space, a point is said to recur if

$$B_r(X_0) = \{X : \|X - X_0\| \leq r\}.$$

The pairwise distances between all the N-amino acids-long subsequent windows (the M rows of EM) are computed, and all the distances smaller than r are scored as recurrent. The application of this computation produces a recurrence plot (RP), that is, a symmetrical $M \times M$ array in which a point is placed at (i, j) whenever a point X_i is close to another point X_j. Graphically this can be indicated by a dot. Thus recurrence plots simply correspond to the distance matrix between the different epochs (rows of EM) filtered—by the action of the radius—to a binary 0/1 matrix, with 1 (dot) for distances falling below the radius and 0 for distances greater than the radius. Figure 3 reports the RPs of two representative phycocyanins. Because graphical representations may be difficult to evaluate, Zbilut and Webber[14] developed several strategies to quantify features of such plots originally pointed out by Eckmann et al.[13] The quantification of recurrences consists of the generation of five variables: (1) REC, percent of plot filled with recurrent points; (2) DET, percent of recurrent points forming diagonal lines, with a minimum of adjacent points equal to the predefined parameter "line"; (3) ENT, Shannon information entropy of the line length distribution; (4) MAXL, length of longest deterministic segment; and (5) TREND, measure of the decreasing rate of recurrent points away from the central diagonal expressed as the slope of the linear function linking identity in time and number of recurrences. These five indices give a summary of the autocorrelation structure of the series. The application of RQA implies the a priori setting of the measurement parameters embedding dimension, radius, and line (the minimum number of adjacent recurrent

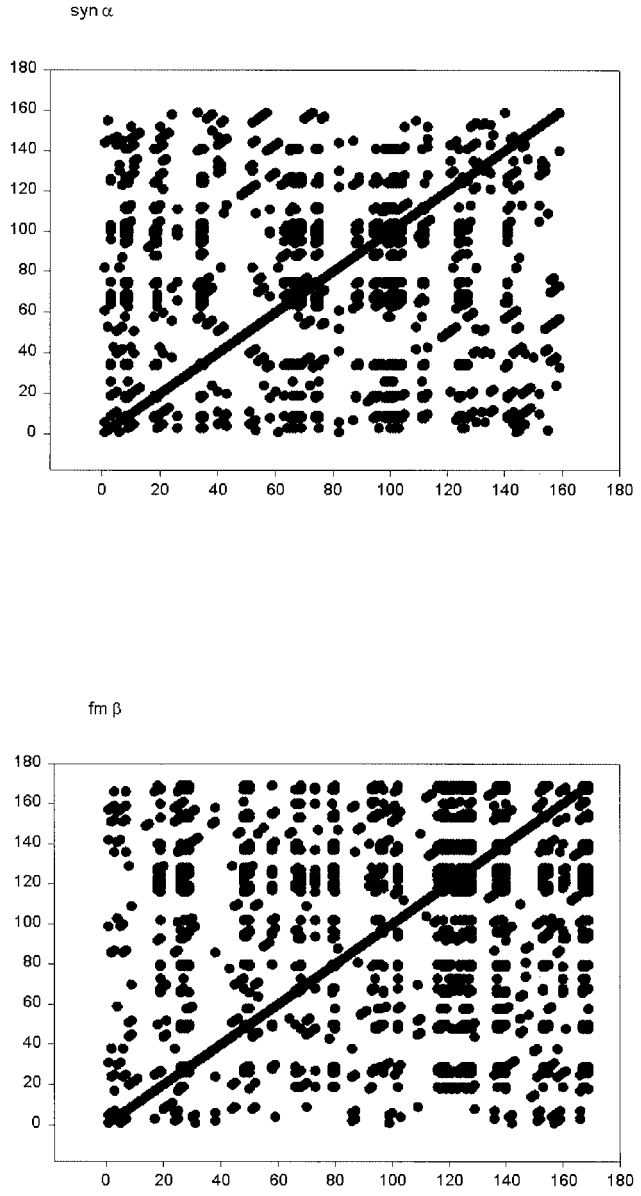

Fig. 3. RQA plots relative to *syn* α-chain and *fm* β-chain.

points to be considered as deterministic). On the basis of our previous studies of protein sequences the above parameters were set to: embedding dimension = 4; radius = 3; and line = 3.

The RQA software can be freely downloaded from: http://homepages.luc.edu/~cwebber/.

SVD. At odds with the more recent RQA, SVD is a well-established method, used for many years in physical as well as in social and biological sciences.[10,12]. Basically, SVD corresponds to principal component analysis (PCA).[17] The term SVD is preferred to the term PCA in physical applications and, more in general, when dealing with dynamic phenomena. As in PCA, the aim of SVD is to project an originally multidimensional phenomenon onto a reduced set of new axes, which are orthogonal to each other and represent the basic modes of the analyzed data.[12] When applied to an originally monodimensional series, SVD requires that the original series is represented on a multidimensional space by the agency of the embedding procedure, so giving rise to an $M \times N$ matrix. A basic theorem of linear algebra is that each $M \times N$ matrix X can be expressed as

$$X = USV^T \tag{1}$$

where the matrices U and V are of dimensions $M \times K$ and $N \times K$, respectively, and fulfill the relations $UT_U = V^T{}_V = 1$. The $K \times K$ matrix S (typically the covariance matrix) is diagonal and has its diagonal elements (singular values) arranged in descending order $s1 > s2 > s3 \ldots > sk > 0$.

In intuitive terms this means that the original data can be projected onto a new set of coordinates US (principal component scores or eigenfunctions) such that no original information is lost; each element of X can be reconstructed by the equation

$$Xij = \Sigma\ Uik\ Sk\ Vjk \tag{2}$$

$$K = 1 \text{ to } N.$$

With the expansion truncated to A terms (with $A < N$) one obtains the summation

$$Xij = \Sigma Uik\ Sk\ Vjk + Eij \tag{2a}$$

$$K = 1 \text{ to } A$$

where the squared error term $\Sigma E^2{}_{ij}$ is a minimum. What differentiates (2) from (2a) is the presence of the error term Eij and the limitation of the summation to a lower number of coordinates with respect to the original data set. The fact that the error term is a minimum implies that the projection of the original data on the new component space spanned by a smaller number of dimensions ($A < N$) is optimal in a least-squares sense. Thus, the meaningful (signal-like) part of the information is retained by the first principal components, whereas the noise is discarded in the error term. In other words, the most correlated portion of information is retained by the first components, while all the singularities are discarded in the minor components. In this work the SVD was applied to an 8D embedding and the first three eigenfunctions (components) were extracted.

Note that we used an 8D embedding for SVD, while RQA was applied to a 4D EM. This difference stems from the fact SVD is more sensitive to periodic patterns spanning the entire sequence (like secondary structures), while RQA is more adapted to the identification of local structures like hydrophobicity singularities potentially involved in folding initiation. This "global versus local" character of the two techniques influences the choice of the analysis window.

The proportion of variance explained by the first eigenvector ($E1$) can be considered as an inverse index of complexity of the series.[10] For this reason, $E1$ was the first SVD-related index used in this work to characterize the protein sequences.

TABLE II. Summary of the Descriptors Used in the Analysis (See Details in the Text)

ENTNUM	Shannon's entropy of amino acid composition
REC	Percentage of recurrence
DET	Percentage of determinism
ENT	Shannon's entropy of deterministic line distribution
MAXL	Maximal length of deterministic lines
TREND	Slope of the relation between distance in time and number of recurrences
AVE	Average hydrophobicity
SD	Standard deviation of hydrophobicity
R	Correlation coefficient between adjacent residues
LZ	Lempel–Ziv complexity of sequence
$E1$	First eigenvalue of SVD-filtered sequence
FD	Dominant Frequency of SVD-filtered sequence
$SP1$	Correlation coefficient of SVD-filtered spectrum with cluster 1 spectra
$SP2$	Correlation coefficient of SVD-filtered spectrum with cluster 2 spectra
$SP3$	Correlation coefficient of SVD-filtered spectrum with cluster 3 spectra
$SP4$	Correlation coefficient of SVD-filtered spectrum with cluster 4 spectra
R_{ABS}	Absolute value of the correlation coefficient between adjacent residues

Following the approach of the Mandell group,[5] after filtering with SVD the hydrophobicity-coded sequences we derived the power spectra of the filtered sequences. These power spectra , expressing the most prominent periodicity patterns (amino acids $^{-1}$ units) entail the relevant information about the regularities of hydrophobicity patterning along the chain, and were demonstrated to be related to the presence of secondary and supersecondary structures,[18] as well as to be able to predict peptide–receptor and protein–protein interactions.

In the present analysis, the results from the spectral analysis were used in two ways to describe the sequences. First, we considered the dominant frequency (FD) of the power spectra. Second, we compared to each other, by a holistic procedure, the whole profiles of digitized spectra. The digitized spectra (1000 points each) were submitted to oblique principal component analysis[7,19] [OPC, VARCLUS procedure in the SAS Statistical Software (SAS Institute Inc., NY, 1990, version 8.0)]. This classified the 32 spectra into 4 clusters on the basis of their relative cross-correlations (the clusters collected the spectra most correlated to each other and thus with a similar shape). For each sequence four coefficients (SP1–SP4) were computed; these coefficients were the Pearson correlation coefficients of each spectrum with each of the four clusters, thus providing a global description of the spectral shape.

The SVD and relative spectra were computed with the software CDA (Chaos Data Analyzer) of the American Physical Society from http://www.aps.org/.

Nonembedding-related methods

The nonembedding related measures we adopted for this work were entropy of the amino acid composition (ENTNUM), average hydrophobicity value (AVE), standard deviation of residues hydrophobicity (SD), LZ complexity (LZ), and Pearson's correlation between adjacent values in both relative and absolute units (R and R_{ABS}) (Table II).

ENTNUM is the Shannon's entropy formula applied to the relative frequency of each amino acid species in the protein. It has no relation with the amino acid order but only with the protein composition. AVE is the average value of hydrophobicity for the given sequence and SD its standard deviation.

The Lempel–Ziv complexity (LZ) is one of the most widely used descriptor for algorithmic complexity because of its easy implementation and wide applicability.[20] LZ is an order-dependent measure and works as follows. First, the hydrophobicity sequence is transformed into a binary format , substituting 1 for the higher-than-median values and 0 otherwise. This binary sequence is then analyzed, trying to generate any subsequent configuration of 1s and 0s from the previous one using the two operators "copy" and "insert" on the initial sequence. Starting from an initial random sequence, Sr, the procedure progressively reconstructs any predefined series. The number of instructions ("copy" plus "insert" operations) needed to produce the series, normalized by the number of instructions needed to generate the corresponding random sequence, is the LZ index. LZ is thus the numerical approximation of the mathematical notion of algorithmic complexity, defined as number of "rules" needed to generate a given series.

Pearson's correlation (R) corresponds to the well-known statistical formula

$$R = \mathrm{Cov}(XY)/\sqrt{\mathrm{Var}(x)*\mathrm{Var}(y)}, \quad (3)$$

where X and Y are adjacent values in the series, Cov is the covariance, and Var is the variance. This is a measure of how strongly the hydrophobicity of an amino acid correlates with the hydrophobicity of its immediate neighbor, so pointing to a special kind of deterministic structure recognized to have structural consequences for proteins.[21,22] Absolute correlation (RABS) was used to evaluate the neighboring amino acid correlation, independently from the sign (i.e., a hydrophobic/hydrophilic pattern gives rise to a negative R, while a hydrophobic/hydrophobic or equivalently hydrophilic/hydrophilic pattern gives rise to positive R values).

TABLE III. Data Set

Name	ENTNUM	REC	DET	ENT	MAXL	Trend	AVE	SD	R	LZ	$E1$	FD	SP1	SP2	SP3	SP4	R_{ABS}
1agla	4.049	3.527	33.183	1.446	8	−14.73	−1.096	4.57	−0.14	1.13	1.47	0.499	0.877	0.280	0.956	0.467	0.14
1agme	4.032	3.38	43.16	1.56	6	−1.51	−1.23	4.63	−0.11	1.18	1.39	0.286	0.912	0.324	0.796	0.643	0.11
1ana	3.987	3.097	36.55	1.456	6	−4.85	−1.18	4.61	−0.18	1.13	1.49	0.499	0.924	0.198	0.954	0.448	0.18
1anac	3.906	5.05	42.36	1.59	6	−18.51	−1.09	4.56	−0.12	1.18	1.37	0.281	0.968	0.393	0.791	0.738	0.12
1cc	4.011	5.34	42.03	1.6	6	−7.61	−0.86	4.48	−0.17	1.13	1.6	0.332	0.867	0.198	0.805	0.466	0.17
1elo	4.018	3.65	40.74	1.67	7	−0.51	−0.97	4.49	−0.22	1.18	1.56	0.499	0.839	0.182	0.987	0.343	0.22
1fm	4.008	5.453	44.96	1.743	6	−10.32	−0.86	4.44	−0.1	1.18	1.19	0.499	0.531	0.0432	0.891	0.024	0.1
1gal	4.008	3.38	29.95	1.04	6	−12.42	−1.111	4.52	−0.13	1.13	1.47	0.275	0.913	0.222	0.698	0.580	0.13
1ms	3.913	3.05	30.29	1.59	6	−4.47	−1.021	4.57	−0.2	1.18	1.49	0.499	0.938	0.194	0.873	0.488	0.2
1pan	4.006	6.053	45.25	1.565	7	−14.47	−0.88	4.52	−0.09	1.09	1.27	0.499	0.409	−0.047	0.784	−0.079	0.09
1por	4.027	4.7	39.49	1.53	6	−28.7	−0.9	4.44	−0.13	1.13	1.37	0.499	0.858	0.221	0.975	0.397	0.13
1rho	4.006	5.5	46.6	1.6	7	−23.37	−0.85	4.37	−0.14	1.18	1.44	0.499	0.740	0.170	0.967	0.252	0.14
1sp	4.033	3.49	36.3	1.37	6	9.012	−1.19	4.68	−0.13	1.087	1.29	0.499	0.973	0.334	0.839	0.659	0.13
1syn	3.983	4.641	38.59	1.754	7	0.765	−0.99	4.53	−0.17	1.13	1.52	0.276	0.988	0.348	0.804	0.699	0.17
1syn63	3.906	5.13	41.46	1.61	6	−19.56	−1.07	4.56	−0.15	1.18	1.55	0.28	0.881	0.328	0.563	0.742	0.15
1syv	4.012	3.61	41.28	1.67	7	0.395	−0.91	4.45	−0.22	1.18	1.55	0.499	0.846	0.191	0.987	0.358	0.22
2agla	3.903	3.76	37.87	1.58	9	−14.3	−1.21	4.87	−0.09	1.12	1.41	0.29	0.664	0.868	0.363	0.970	0.09
2agme	3.971	4.02	43.26	1.62	11	30.99	1.19	5.01	0	1.16	1.43	0.283	0.914	0.537	0.657	0.848	0
2ana	3.963	3.46	35.01	1.328	9	−12.32	−1.35	4.96	0.04	1.03	1.44	0.285	0.689	0.656	0.263	0.963	0.04
2anac	3.888	4.19	37.98	1.39	8	−12.83	−1.3	5.13	0.003	1.12	1.39	0.13	0.650	0.873	0.374	0.935	0
2cc	3.908	4.21	38.36	1.81	8	−14.13	−1.15	4.9	−0.02	1.22	1.4	0.285	0.763	0.736	0.455	0.930	0.02
2elo	3.915	3.74	32.96	1.48	9	−20.37	−1.23	4.95	0.003	1.12	1.43	0.286	0.673	0.789	0.277	0.996	0
2fm	3.893	5.375	39.71	1.549	11	3.473	−0.98	4.84	0.01	1.12	1.32	0.128	0.398	0.983	0.326	0.765	0.01
2gal	3.962	4.53	37.64	1.62	9	−20.3	−1.15	4.96	−0.04	1.12	1.43	0.289	0.599	0.890	0.263	0.979	0.04
2ms	3.955	3.29	35.97	1.2	9	−9.39	−1.38	4.96	−0.04	1.16	1.46	0.283	0.712	0.602	0.271	0.944	0.04
2pan	3.845	4.994	35.68	1.441	12	−4.019	−0.96	4.82	0.02	1.16	1.39	0.129	0.201	0.951	0.052	0.702	0.02
2por	3.926	4.06	28.25	1.5	8	−0.36	−1.17	4.85	0.017	1.12	1.34	0.123	0.404	0.954	0.347	0.752	0.017
2rho	3.846	5.13	41.56	1.59	9	−1.76	−1.04	4.79	0.002	1.08	1.31	0.121	0.324	0.953	0.194	0.760	0
2sp	3.944	5.523	37.117	1.48	12	−23.4	−0.93	4.82	0.03	1.036	1.41	0.128	0.293	0.981	0.089	0.805	0.03
2syn	3.923	4.663	31.87	1.565	11	−17.87	−1.18	5.02	0.07	1.16	1.33	0.294	0.375	0.963	0.126	0.868	0.07
2syn63	3.883	4.19	35.63	1.36	9	−10.59	−1.23	5.07	−0.01	1.18	1.38	0.499	0.741	0.381	0.835	0.472	0.01
2syv	3.915	3.74	32.96	1.48	9	−20.37	−1.18	4.91	0.04	1.12	1.43	0.286	0.673	0.789	0.277	0.996	0.04

Prefix 1, α-chain; prefix 2,- β-chain.

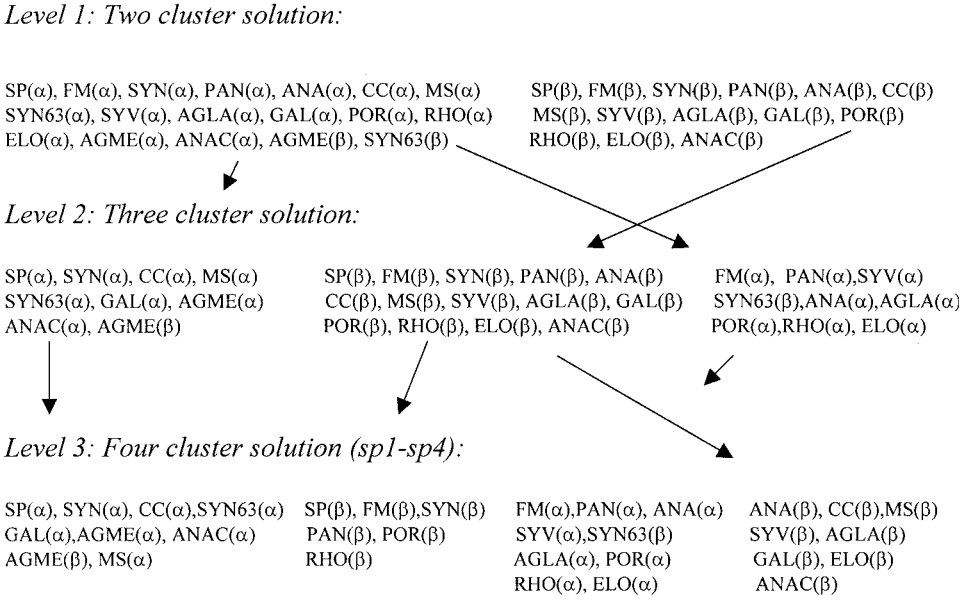

Fig. 4. Clusterization of power spectra of SVD-filtered sequences, according to oblique principal component analysis (VARCLUS procedure). The clusters collect spectra highly intercorrelated and as independent as possible from the spectra of the other clusters. In VARCLUS, the formation of new clusters ends when a threshold of 0.60 of between-clusters-correlation is reached.

RESULTS AND DISCUSSION

The sequences in Table I were coded in terms of the Schneider and Wrede hydrophobicity scales[11] and then subjected to a range of signal analysis techniques (see Methods section) that generated 17 descriptors for each sequence. The descriptor values are reported in Table III. The two main "suites" of descriptors were based on RQA, which points to hydrophobicity local motifs, and SVD, which detects global periodicities in the hydrophobicity distribution. Thus, Table III represented the starting material for the elucidation of the relation linking α and β subunits at the level of sequence-based information.

Sequence Homology

As stated in the Introduction, the pairs of interacting subunits do not show any noticeable sequence homology. Figure 2 displays the sequence homology tree (http://clustalw.genome.ad.jp) of the phycocyanins listed in Table I. The α and β subunits are clearly separated into two main branches of the tree.

It is interesting that the above partition into α and β branches was mirrored by the shapes of the hydrophobicity power spectra obtained via SVD. The SVD-filtered spectra of the primary structures were subjected to a VARCLUS clustering procedure: The analysis generated a hierarchical clustering whose first division pointed to an almost complete separation of the α and β subunits into distinct clusters (Fig. 4). This separation remained also at further partitions that pointed to progressively minor differences among the spectra. Figure 5 displays the four main types of spectra, corresponding to the centers of the four SP1–SP4 clusters.

Protein sequences are quasirandom strings[23]; this implies that even minor modifications of the protein sequences can give rise to relevant modifications of the spectra.[7,24] On the contrary, we observed a tight clusterization of spectra into a few basic modes; this points to the existence of strict structural constraints that shape the observed hydrophobicity distribution periodicity. As a consequence, the leading modes of the spectra evidenced in Figure 5 can be considered the "hydrophobicity distribution counterparts" of the structural requirements that rule the generation of correctly shaped αβ complexes.

Characterization of the Sequences

After the generation of 17 descriptors for each sequence (Table III) a major analytic step was the application of principal component analysis (PCA)[17] to the Table III data. The original 17D space was reduced to a 3D space [principal components (PCs)] maintaining a large part of the original information. The components are by construction orthogonal to each other: Each represents an independent aspect of the autocorrelation structure of the protein sequences. Each statistical unit (here, protein sequence) has values (component scores) on the PCs, which are the new variables that describe the data; these scores replace the values relative to the original 17 variables and can be used in the subsequent analyses. Moreover, this reduction in dimensionality rules out the risk of finding chance correlations in the analysis of the data.[25]

PCA gave rise to the eigenvector distribution reported in Figure 6: A leading first PC, explaining alone 46% of total variability, is followed by relatively minor components. The second PC explains 17% of the variance and the third

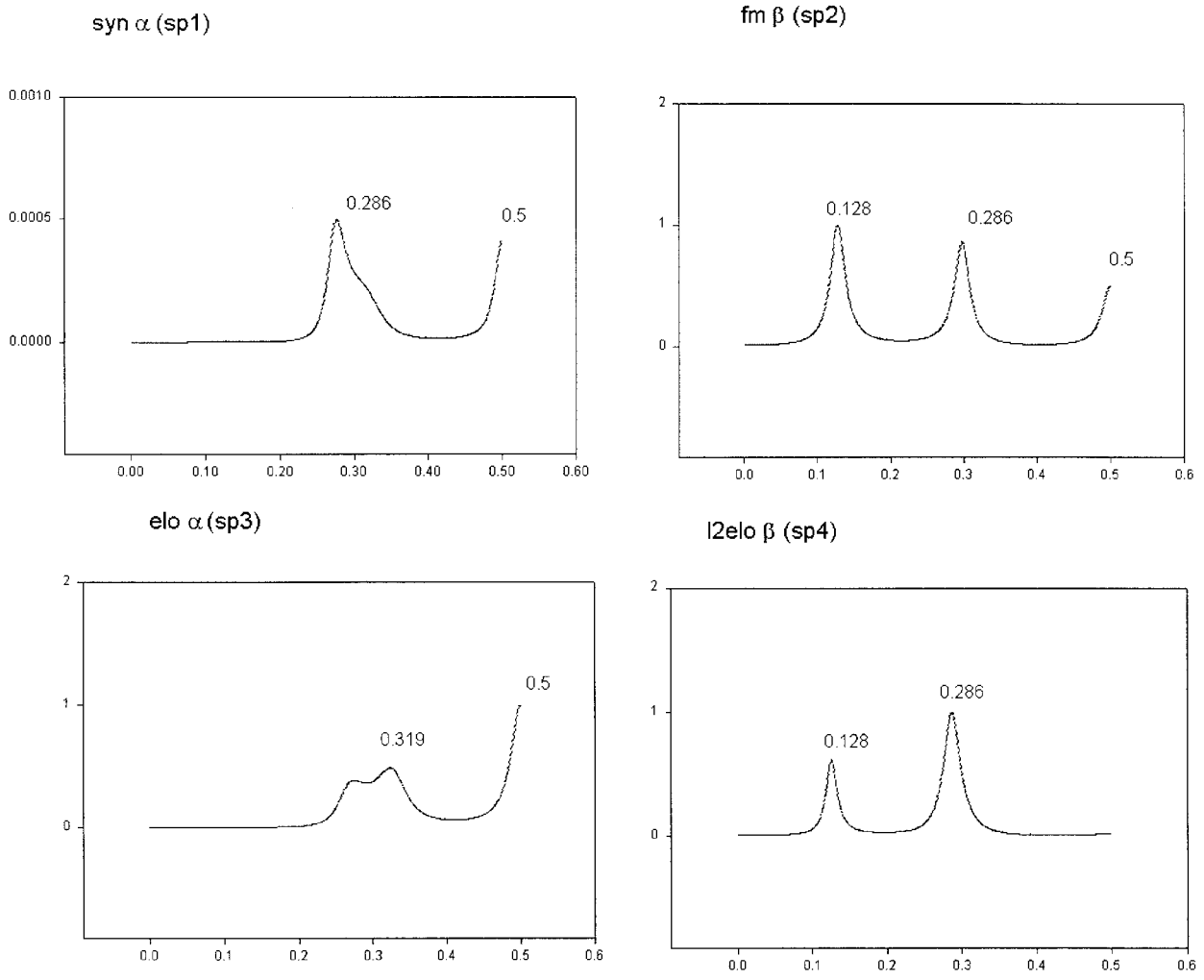

Fig. 5. Spectra most correlated with the centers of each cluster. The X-axis is expressed as aa^{-1}; the Y-axis is an adimensional measure of power. The frequency values of the peaks are reported.

PC explains 8%. The PCs after the third one are barely distinguishable from noise. Thus, a three-component solution, collectively explaining 72% of total variance, was chosen for summarizing the data.

Table IV reports the component loadings, that is, the correlation coefficients between the original variables and the extracted components. These loadings allowed us to attach an explanation to the component-based representation. The variables more heavily loaded on PC1 are ENTNUM ($r = 0.72$), MAXL ($r = -0.82$), SD ($r = -0.88$), R ($r = -0.93$), FD ($r = 0.80$), SP2 ($r = -0.94$), SP3 ($r = -0.94$), SP4 ($r = -0.77$), and RABS ($r = 0.89$). From a purely compositional (not order-dependent) viewpoint, high values of PC1 correspond to a varied amino acid composition (positive correlation with ENTNUM), together with relatively homogeneous hydrophobicity (negative relation with SD). From an order-dependent perspective, high values of PC1 correspond to an alternate hydrophobic–hydrophilic pattern (as evidenced by R and R_{ABS} loadings and the high positive loading of FD pointing to this high-frequency periodicity), short-range ordering (negative relation with MAXL), and an SP3 − SP1-like pattern, as opposite to SP2 − SP4 (Fig. 5). As shown in Figure 7 (PC1 − PC2 space), this profile corresponds to the α character of the sequences (high PC1 values), as opposed to the β character. As a matter of fact, there is a perfect separation between the α and β structures in the PC1/PC2 plane, where all the β sequences have PC1 values uniformly lower than the mean and the α sequences have values higher than the mean. This neat separation between α and β subunits overperforms the imperfect separation obtained by sequence alignment. In fact, in the principal component space, even the α and β chains of *syv* are correctly discriminated. Note that the "consensus" information of the different hydrophobicity autocorrelation descriptors as summarized by PC1 score recovers the basic α versus β opposition, clearly evidenced by sequence alignment, without the need of any alignment between different sequences, but simply comparing their general autocorrelation features. This implies the possibility of sensible comparisons of nonhomolog sequences that is precluded by classic sequence alignment strategies.

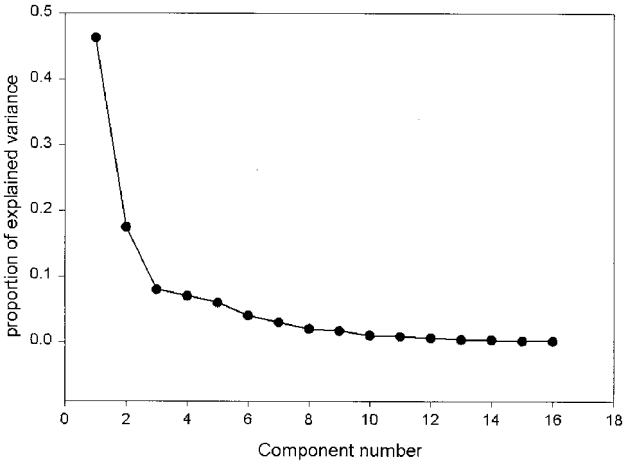

Fig. 6. Eigenvalues distribution of the PCA applied to Table III data. The figure shows that the three-components solution emerges from the noise floor.

TABLE IV. Principal Component Analysis Solution of Table III Data

	PC1	PC2	PC3
ENTNUM	0.724	−0.121	−0.328
REC	−0.054	0.927	0.016
DET	0.408	0.626	0.232
ENT	0.269	0.537	0.584
MAXL	−0.816	0.195	0.050
Trend	0.160	−0.156	−0.240
AVE	0.519	0.720	−0.021
SD	−0.882	−0.240	0.097
R	−0.934	0.141	−0.102
TZ	0.390	−0.018	0.630
E1	0.326	−0.466	0.459
FD	0.796	−0.096	−0.162
SP1	0.661	−0.578	0.264
SP2	−0.940	0.021	0.141
SP3	0.942	−0.073	−0.033
SP4	−0.772	−0.412	0.341
R_{ABS}	0.891	−0.159	0.061
Expl. var. (%)	46.3	17.5	8.4

The component loadings (i.e., correlation coefficients of the original descriptors with the Components) are reported, together with the percentage of explained variation (Expl. Var.) for each component.

Inspection of Figure 7 also shows that the elements of each αβ pair have barely equivalent positions in PC2 and occupy almost symmetrical positions in the PC1,PC2 plane. PC2 is mostly linked to the amount of recurrence of the sequences (REC has a loading of 0.93 with PC2, Table IV). The α and β structures display a practically coincident value of recurrence (average recurrence equal to 4.31 and 4.30 for the α and β sets, respectively) and consequently of PC2. PC2 is thus a sequence-based descriptor catching a feature that is conserved across α and β structures, while varying across different organisms.

Correlations Between Interacting Pairs

A next and final step of the analysis was to check, in a quantitative way, the relatedness of the α and β subunits, within each pair, based on the PC values. As stated above, the PCs are summary descriptions of the set of 17 descriptors generated for each sequence. Canonical correlation analysis[17] was applied to the data field spanned by the first three PCs.

At odds with PCA, which implies a symmetrical character of all the variables spanning the data set, canonical correlation is based on the existence of two separate sets of variables (X and Y sets) defining the statistical units. Canonical correlation finds the two linear combinations of respectively X and Y variables whose mutual correlation is a maximum; these two linear combinations constitute the first canonical variates pair. After this step, the analysis looks for other canonical variates pairs, orthogonal to the first ones, that explain progressively lower amounts of correlation between the two X and Y sets. In analogy with principal components, whose meaning is interpreted by means of the correlation coefficients (loadings) of the component with the original variables, even the canonical variates are interpreted in terms of their correlation with the elements of the X and Y sets. In summary, canonical correlation can be considered an extension to the multidimensional spaces of the ordinary concept of correlation in bivariate situations. In this case, the two sets of variables

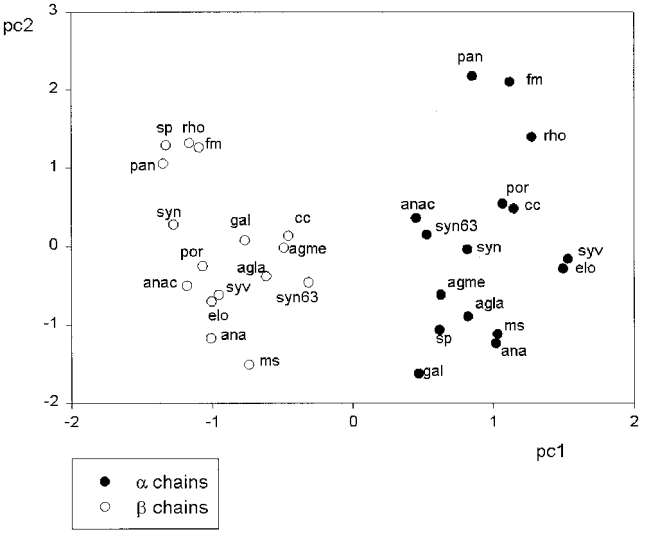

Fig. 7. Projection of proteins on the PC1/PC2 plane. The α and β chains populations are clearly separated in the plane.

to be compared by canonical correlation were PC1–PC3 of, respectively, α and β subunits relative to the same organism. In practice, canonical correlation was applied to a matrix having (a) the organisms in the rows acting as statistical units and (b) the three PCs relative to the α units and the three PCs relative to the β units in the columns acting as X and Y variables sets.

The first canonical variates of the α versus β subunits reached an extremely high, statistically significant Pearson correlation coefficient: $r = 0.89$ ($P < 0.0001$). Moreover, in the space spanned by the canonical variates relative to α and β subunits there was an almost linear

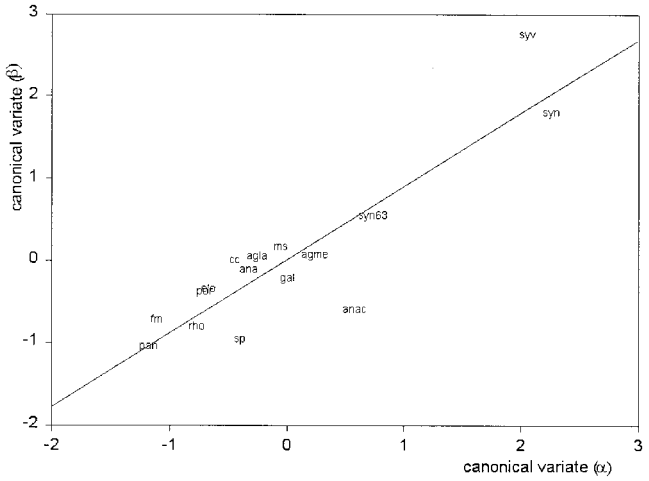

Fig. 8. Canonical correlation between α and β chains of different organisms. It scores a correlation $r = 0.89$ and represents the image in light of protein–protein interaction.

disposition of the different organisms (Fig. 8). The first canonical variates pair was almost exclusively based on PC1 that scored a correlation coefficient of around 0.9 for both α and β sets. This implies that the between-subunits interaction can be modeled by only one variable (PC1); as a matter of fact the direct correlation between PC1 scores of the α and β subunits of the same organisms reaches a Pearson correlation coefficient equal to 0.84.

It should be remarked that the entire process (coding of the sequences; subsequent calculation of dynamic descriptors, then of PCs and canonical variates) can be applied to newly sequenced (or artificially designed) α and β subunits, and their efficiency in forming complexes can be theoretically predicted.

The canonical variable is thus a recipe, expressed in terms of pure sequence information, for generating an interacting pair or, in other words, a way for judging of the strength of interaction of two protein sequences.

CONCLUSIONS

In analogy with the intramolecular interactions that shape the 3D structure of the monomeric proteins, also the protein–protein interactions that give rise to multimeric complexes such as the phycobilisomes are driven by the formation of weak molecular bonds in the range $\Delta G = 2$–7 kcal/mol. The kinetic energy of the background of heat-generated molecular motion is of the order of $\Delta G = 0.6$–1.0 kcal/mol at 25° C.[9] This implies that, to survive, any relatively stable interaction must be cooperative, that is, must be based on the simultaneous generation of a multiplicity of relatively weak bonds.[6] The dominant character of the hydrophobic interaction in the realm of intermolecular forces traces back this concept to the need of a specific hydrophobicity distribution along the chain to support a given interaction.[9] This is the chemical physical basis of the work of Mandell's group in the search for "mode matches" in the hydrophobicity distribution spectra of interacting proteins,[4,5] as well as of our results relative to the correlation between the hydrophobicity recurrence of interacting pairs.[8] It is important to stress that this "coherence" between interacting pairs does not necessarily imply neither a structural nor a sequence resemblance between them but, more properly, a sort of "reciprocal fitting" of the two (mainly hydrophobical) potentials. The results reported here provide further support to the above notions and clearly indicate that a comprehensive description of hydrophobicity distribution along the sequences can efficiently model protein–protein interactions. Moreover, the application of the present model provides a practical way for predicting still unobserved protein–protein interactions.

An immediate question is, "how can these results be interpreted in classic 3D structure terms?" Mandell and colleagues interpreted the peaks of the hydrophobicity distribution spectra (periodicity) in terms of secondary and supersecondary structures[5] (see also [18]). Following their interpretation, we observed (Fig. 5) the presence of two peaks related to the content in α helices, namely, those around 0.3 aa^{-1} (frequencies 0.286 and 0.319 in Fig. 5, correspondent to a period of 3.5 and 3.13 amino acids, respectively). The periodicity of the two peaks is close to the typical α helix pace (around 3.6 amino acids) and is consistent with the prevalently α helix character of the phycocyanins. The most obvious explanation is that the α and β subunits interact via contact of their α helices, being the α helix-related periodicity the most prominent common peak of the two partners. But, it should also be remembered that the two subunits are almost totally arranged into α helices and thus this conclusion lacks of any specificity; it is almost trivial to state that two mainly α helical structures sharing large contact interfaces are connected by their α helices.

An intriguing feature of the αβ interaction is the fact that the motion of one of the subunits is strictly related to the motion of the other, as demonstrated by the normal mode analysis by Kikuchi et al.[26] This implies a dynamic link between the two subunits, involving the transfer of the relative motion probably at the level of the maximal flexibility zones.

We are beginning to appreciate that many of the crucial zones for protein–protein interactions correspond to natively unfolded zones of the protein structures, that is, highly flexible portions of the 3D architecture.[27] According to our unpublished results on a random sample of Swiss–Prot repository made of 1141 protein sequences, these zones correspond to highly deterministic, recurrent portions of the proteins (as far as hydrophobicity distribution along the chain is concerned). The above perspective may be an explanation for the hydrophobic–hydrophilic patterning of amino acids, as measured by both R and the high-frequency (0.5 correspondent to 2 amino acids) peak of the spectra (Fig. 5) as a supplementary structural counterpart of phycocyanin subunits interaction. In fact, the presence of strong adjacent residues correlation lowers the relative complexity of the sequence. These low-complexity zones, according to Dunker's data,[27] may corre-

spond to partially unfolded patches involved in protein–protein interactions.

The presence of a low-frequency peak (0.128 aa^{-1}, correspondent to 7.81 amino acids) specific for the β sequences is of difficult interpretation (Fig. 5). On a purely mathematical perspective, it may be a "harmonic" of the fundamental α helix rhythm (7.81 is approximately the double of the 3.6 periodicity of the α helix). However, this interpretation has no simple support in what we know about the crystalline structure of the phycocyanins, which do not show any marked difference between the α and β subunits able to explain the pure β character of this periodicity.

It should be added that it is not possible to go from the hydrophobicity patterning along the chain directly to 3D structure: the case of α and β subunits of phycocyanins, which couple a marked homogeneity in 3D structure with an heterogeneity in sequence and hydrophobicity patterning, is a clear example of this reality. More reasonably, we can hypothesize that the level of hydrophobicity patterning along the chain is at an intermediate level between sequence and structure, thus encompassing all those "dynamic" properties of the protein behavior not evidenced by the pure crystalline 3D arrangement. Pure sequence, on one hand, and pure structural information, on the other hand, are inadequate to give reason for the specific interaction between phycocyanin subunits. As a matter of fact, the pure sequence information splits the α and β subunits into two disjoint clusters (Fig. 2), with no evidence of coupling of interacting pairs. At the same time, structural information does not provide evidence of any basic difference between the α and β subunits, implying the necessity of the heterodimer as the basic brick of phycobilisome. On the contrary, the intermediate level of hydrophobicity patterning was the only one able to give a quantitative expression to the interaction between α and β subunits.

What still lacks an explanation is how we can discriminate, in the case of two interacting partners, the correlations due to the sharing of common structural features (like in this case the presence of large portion of α helices giving rise to peculiar peaks in the SVD spectrum) from the correlations directly involved in the interaction process. This is still an open problem that waits for a general solution, even if the work of Dunker's group[27] sketching a link between interaction hot spots and natively unfolded structures is probably indicating a solution. On a more operational ground, the link between sequence features correlation and interaction of two protein systems was elegantly explained by Cohen's group in coevolutionary terms.[28,29] Two proteins that interact share some mutual structural constraints that limit the possibility of "accepted" random genetic independent variation of the two systems imposing a correlation to the mutational spectra of the two interacting pairs. This correlation is apparent in terms of correlation between the phylogenetic trees built upon the two interacting pairs. As a matter of fact Cohen's group demonstrated[28,29] how the phylogenetic trees coming from two interacting proteins are more correlated than the phylogenetic trees coming from two noninteracting systems. The above approach is based on pure sequence alignment and is constrained by the need to have a consistent random genetic drift decorrelating the phylogenetic trees of noninteracting pairs. In previous work[8] we demonstrated our method worked well in a situation of lack of genetic random drift (two viral proteins) in which Cohen's approach failed to discriminate between interacting and noninteracting pairs. Moreover, our method, being based on chemicophysical properties of the interacting system, allows for a mechanistic interpretation of the interaction. On the other side, Cohen's approach is surely more powerful than ours when looking for unexpected interactions between protein systems not previously known to interact. In conclusion, both methods work along the same line of reasoning (i.e., interaction implies the existence of mutual constraints between the two systems), but while our method is more efficient in the fine-tuning of the interaction process (i.e., how to modify the sequence of the partners so to improve interaction) Cohen's approach is more efficient to explore massive genomic data and chasing for possible interaction partners.

AKNOWLEDGMENTS

The continued interest of Marco Crescenzi is gratefully acknowledged. Work partially supported by NSF/NIH grant no. 0240 230.

REFERENCES

1. Schirmer T, Bode W, Huber R. Refined three-dimensional structure of two cyanobacterial C-phycocyanins at 2.1 and 2.5 Angstrom resolution. J Mol Biol 1987;196:677–695.
2. Glazer AN. Phycobilisomes: structure and dynamics. Annu Rev Microbiol 1982;36:173–198.
3. Adir N, Dobrovetsky Y, Lerner N. Structure of C-phycocyanin from the thermophilic cyanobacterium synechococcus vulcanus at 2.5 Angstroms: structural implications for thermal stability in phycobilisome assembly. J Mol Biol 2001;313:71–81.
4. Mandell AJ, Selz KA, Shlesinger MF. Predicting peptide–receptor, peptide–protein and chaperone–protein binding using patterns in amino acid hydrophobic free energy sequences. J Phys Chem B 2000;104:3953–3959.
5. Mandell AJ, Owens MJ, Selz KA, Morgan WN, Shlesinger MF, Nemeroff CB. Mode matches in hydrophobic free energy eigenfunctions predict peptide–protein interactions. Biopolymers 1998;46: 89–101.
6. Dobson CM, Karplus M. The fundamentals of protein folding: bringing together theory and experiment. Curr Opin Struct Biol 1999;9:92–101.
7. Giuliani A, Benigni R, Zbilut JP, Webber CL Jr, Sirabella P, Colosimo A. Nonlinear signal analysis methods in the elucidation of protein sequence structure relationships. Chem Rev 2002;102: 1471–1491.
8. Giuliani A, Tomasi M. Recurrence quantification analysis reveals interaction patterns in paramyxoviridae envelope glycoproteins. Proteins 2002;46:171–176.
9. Mandell AJ, Selz KA, Owens MJ, Shlesinger MF, Gutman DA, Arcuragi V. Hydrophobic mode-targeted, algorithmically designed peptide ligands as modulators of protein thermodynamic structure and function. In: Raffa RB, editor. Drug–receptor thermodynamics: introduction and applications. New York: Wiley & Sons; 2001. p 655–700.
10. Giuliani A, Colafranceschi M, Webber CL Jr, Zbilut JP. A complexity score derived from principal component analysis of nonlinear order measures. Physica A 2001;301:567–588.
11. Schneider G, Wrede P. Artificial neural networks for computer-based molecular design. Progr Biophys Mol Biol 1998;70:175–222.
12. Broomhead DS, King GP. Extracting qualitative dynamics from experimental data. Physica D 1986;20:217–236.

13. Eckmann JP, Kamporst SO, Ruelle D. Recurrence plots of dynamical systems. Europhys Lett 1987;4:973–977.
14. Webber CL Jr, Zbilut JP. Dynamical assessment of physiological systems and states using recurrence plot strategies. J Appl Physiol 1994;76:965–973.
15. Manetti C, Ceruso MA, Giuliani A, Webber CL Jr, Zbilut JP. Recurrence quantification analysis as a tool for characterization of molecular dynamics simulations. Phys Rev E 1999;59:992–998.
16. Rustici M, Caravati C, Petretto E, Branca M, Marchettini M. Transition scenarios during the evolution of the Belousov–Zhabotinsky reaction in an unstirred batch reactor. J Phys Chem A 1999;103:6564–6570.
17. Lebart L, Morineau A, Warwick KM. Multivariate descriptive statistical analysis. New York: Wiley; 1984.
18. Murray KB, Gorse D, Thornton JM. Wavelet transforms for the characterization and detection of repeating motifs. J Mol Biol 2002;316:341–363.
19. Harman H. Modern factor analysis. Chicago: Chicago University Press; 1976.
20. Kaspar F, Schuster KG. Easily calculable measure for the complexity of spatio-temporal patterns. Phys Rev A 1987;36:842–847.
21. Wang W, Hecht MH. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. Proc Natl Acad Sci USA 2002;99:2760–2765.
22. Baltzer L, Nilsson H, Nilsson J. De novo design of proteins—What are the rules? Chem Rev 2001;101:3153–3163.
23. Weiss O, Jimenez-Montano MA, Herzel H. Information content of protein sequences. J Theor Biol 2000;206:379–386.
24. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 1968;21:170–201.
25. Topliss JG, Edwards RP. Chance factors in studies of quantitative structure–activity relationships. J Med Chem 1979;22:1238–1244.
26. Kikuchi H, Wako H, Yura K, Go M, Mimuro M. Significance of a two-domain structure in subunits of phycobiliproteins revealed by the normal mode analysis. Biophys J 2000;79:1587–1600.
27. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002;41:6573–6582.
28. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Co-evolution of proteins with their interaction partners. J Mol Biol 2000; 299:283–293.
29. Goh CS, Cohen FE. Co-evolutionary analysis reveals insights into protein–protein interactions. J Mol Biol 2002;324:177–192.