Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 6637–6641, September 1988
Biochemistry

Distinctive charge configurations in proteins of the Epstein–Barr virus and possible functions

(charge clusters/periodic charge patterns/sequence repeats/latent genes/first lytic genes)

B. EDWIN BLAISDELL[†] AND SAMUEL KARLIN[‡]

[†]Linus Pauling Institute of Science and Medicine, 440 Page Mill Road, Palo Alto, CA 94306; and [‡]Department of Mathematics, Stanford University, Stanford, CA 94305

Contributed by Samuel Karlin, May 23, 1988

ABSTRACT **The protein products of several open reading frames (ORFs) of the Epstein–Barr virus (EBV) are remarkable in their distribution of charged residues. The nuclear antigen proteins EBNA1–EBNA4 of the EBV latent state contain separate significant clusters of charge of each sign. They (excepting EBNA4) also feature distinctive periodic charge patterns [e.g., $(+, O)_8$, $(O, -, -)_7$] and significant tandem repeats. None of the other ORFs (about 80) of the genome possess the conjunction of these properties. Only the protein encoded from BMLF1, the first immediate early transactivator protein, contains significant multiple charge clusters and periodic charge patterns. All proteins that contain significant repeats also contain at least one significant charge cluster of a single sign. These include EBNA5 and LYDMA produced during latency and BZLF1, whose expression terminates latency and initiates productive growth. It is reasonable to conclude that these aggregate significant charge configurations and repeats are important functionally for the latent existence and for the initiation of the lytic cycle and may be characteristic of these conditions. We discuss how large multimeric protein structures bound together by clusters of unlike charge may provide a mechanism for regulation of the expression of these proteins.**

Epstein–Barr virus (EBV), of the human herpesvirus family, causes infectious mononucleosis and is associated with Burkitt lymphoma and nasopharyngeal carcinoma (1). The entire 172,282-base-pair (bp) EBV genome of one strain (B95-8) has been sequenced by Baer *et al.* (2). This strain immortalizes B lymphocytes, generally resides latently as a replicating episome with restricted copy number, and only rarely enters a productive viral cycle.

In a recent study of the evolution of the 21 × 30 tandem repeats of the oriP region of EBV (3) we noted that the EBV-encoded nuclear antigen EBNA1 contains multiple clusters of charged amino acids of positive sign and of negative sign. Subsequently we found that EBNA1 contains several significant charge patterns such as the iterated forms $(+, O)_n$ and $(O, -, -)_n$ where + designates a positively charged amino acid, − designates a negatively charged amino acid, and O designates an uncharged amino acid. Corresponding charge configurations occur in EBNA2 and EBNA3, the other major nuclear antigens of the latent state. The conjunction of these significant charge sequences prompted the present analysis of the distribution of charged residues in all of the 84 substantial open reading frames (ORFs) in EBV.

Our methods center on the identification of long uninterrupted runs of charged residues, of charge clusters (a 30- to 50-residue segment with an unusually high specific charge content), and of periodic patterns of charge. Only 14 of the 84 ORFs of EBV contain distinctive single charge sequence features. The results are given in Table 1 and displayed in Fig. 1. Possible functions and mechanisms of these charge features are considered in the *Discussion*.

Distinctive charge configurations in a protein may contribute to function and structure in diverse ways. For example, runs of positively or negatively charged amino acids might be expected to be stretched out and exposed structurally. Charge patterns of period two in a β-strand present a straight line of charge on one side of the strand and those of period three in an α-helix present an almost linear curve of charge. It is likely that clusters of charge in the primary amino acid sequence will produce local concentrations of charge on the surface of the tertiary or quaternary protein structure. A general function of clusters of charge may be to establish and stabilize protein conformation. Charge clusters of mixed type increase the stability of the protein in water. The occurrence of multiple charge clusters within one protein might facilitate intramolecular folding and cooperative protein–protein and protein–nucleic acid interactions. Charge clusters and runs appear to be important with respect to protein transport, localization, and regulatory function (4–7).

METHODS

An important methodological problem pertains to the assessment of when an observed charge concentration or pattern is unusual rather than a result of chance fluctuation. This section presents criteria that assess statistical significance of charge configurations (clusters, runs, and periodic patterns).

Concentrations of charge (clusters) are found by computer analysis using a program that scans the sequence of amino acids from beginning to end using a variable window size of lengths 30 up to 50 and counts the number of features (+ charges, − charges, either charge, or no charge) in each window. Windows are sufficiently long to reduce the chance of picking up chance fluctuations and to increase the accuracy of using appropriate distribution theory in assessing the significance of the observed number of features, C, calculated for each window. For an independence (random) model the binomial expected number of charged residues in a window equals Wp and its variance is $Wp(1 - p)$, where p is the fraction in the whole gene. Let $k = |C - Wp|/\sqrt{Wp(1 - p)}$. The probability of finding a value as large as or larger than C in a window is, in the normal approximation, about 0.0001 if $k \geq 3.7$ and therefore is bounded by 0.01 in a set of 100 windows—that is, in a sequence of about 500 amino acids. The factor of 100 is introduced using the conservative Bonferroni inequality to adjust for the many windows examined. We report as significant clusters only regions of length from 30 to 50 residues and $k > 4$ and for any protein exceeding 1000 residues for $k \geq 4.5$. In testing for a positive charge cluster, C is taken

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: EBV, Epstein–Barr virus; EBNA, EBV nuclear antigen; ORF, open reading frame; VZV, varicella-zoster virus.

as $C_+ - C_-$ (C_+ = the count of + charge residues in the window; C_- = the − charge count). A negative charge cluster is based on a significant excess of $C = C_- - C_+$. If neither a positive nor negative cluster is found, we test for a charge cluster of mixed type based on $C = C_+ + C_-$. These are conservative specifications. A check of the procedure on 100 random permutations, on each of 18 genes with balanced charge, excess of positive, or excess of negative (1800 permutations overall), gave results that were in close agreement with expectation.

Significantly long runs of a single letter or pattern of letters in any alphabet are assessed by the theory of Karlin and Ost (8). For example, in a random model, the probability of observing in a sequence of length N a run of length $L = \ln N/(-\ln\lambda) + x$ of a symbol σ, where σ is either "+," "−," or "O" and λ is the frequency of σ, is asymptotically bounded above by $1 - \exp[-(1 - \lambda)\lambda^x]$. We establish the minimum length L^* for a run that gives significance at the 1% level by means of the determination of x that solves $1 - \exp[-(1 - \lambda)\lambda^x] = 0.01$. By similar means we establish significant levels for repetitive patterns of period 2 or 3 (8).

ANALYSIS AND DISCUSSION

For convenience we provide a short review of current knowledge about EBV latency. In the latent state <10% of the viral genome is expressed as stable cytoplasmic poly(A) mRNA, including EBNA1–EBNA4, EBNA6, and LYDMA. The production of EBNA1 (BKRF1) is essential to the replication of plasmids containing the oriP region of EBV (9, 10). EBNA1 protein also transactivates the general capacity of the 21 × 30 bp repeat region of oriP for enhancing the transcription of DNA with various promoters (11). EBNA2 (BYRF1) has been proposed as necessary for growth transformation (12). A 3.5-kilobase-pair (kbp) cDNA composed of several exons processed from a transcription unit of at least 84 kbp is described in refs. 13 and 14 with the most lengthy exon associated mostly with the ORF BERF1 designated EBNA3. Proteins EBNA4 and EBNA6 have been found to be encoded by ORFs BERF2b and BERF4 (15). Protein EBNA5 contains varying numbers of pairs of 22- and 44-residue segments encoded in the multiple copies of the 3.1-kbp repeats (16). The membrane antigen LYDMA (BNLF1) is generally the most abundant transcript produced in the latent state (17). Two ORFs, BZLF1 and BMLF1, are intimately associated with the activation of the productive cycle in normally latent cell lines of EBV. The gene product from BZLF1 disrupts latency and initiates production of the BMLF1 protein that is the first and most abundant protein of the lytic cycle (18, 19). It transactivates its own production and then induces transcription of many other lytic genes (20, 21).

Table 1 (and Fig. 1) report all statistically significant clusters, runs, and periodic patterns of charged residues in the known and putative polypeptides of the EBV.

It is important to distinguish the presence of local charge clusters from global net charge. The histones have high global net positive charge but no clusters, whereas the EBNA proteins with substantially less global net charge have many significant clusters.

Charge configurations in proteins can be characterized depending on sign (positive, negative, mixed), multiplicity (numbers of charge forms), pattern (periodic, irregular), location (quartile of protein sequence), and specific amino acid composition (variation in the use of a charge type).

(*i*) *Proteins with multiple charge clusters of each sign.* This attribute distinguishes EBNA1. There are four charge clusters of a single sign. These are, in order, positive, positive, negative, and negative (see Fig. 1). The second positive and first negative are adjacent. The first and second positives are separated by an extensive (238 residues) uncharged run composed exclusively of glycine and alanine.

The first positive charge cluster contains a $(RG)_5$ iteration. The second positive charge cluster, which has the pattern $(R, O)_{15}$ with a few mismatches, also contains a preponderance of RG doublets. The second positive charge cluster contains a significant DNA repeat region. The carboxyl-terminal segment of EBNA1 presents the pattern $(O, -, -)_7$ including the word DGDEG twice.

Milman and Hwang (22) have reported that the rate of binding of the carboxyl third of EBNA1 to monomer, dimer, and trimer oligonucleotides, constructed to be similar to the repeat palindromes of oriP, is first order in DNA concentration and highly dependent on EBNA1 concentration. A change of EBNA1 concentration from 2.1 to 2.6 μg/ml increases the rate about 60-fold, which implies an n-mer of average $n \approx 20$. This last observation implies a ready formation of large EBNA1 polymers that can be three-dimensional since there are more than two possibly interacting charge clusters per molecule.

(*ii*) *Proteins with at least one charge cluster of each sign.* It is intriguing that the major nuclear latently expressed genes EBNA1–EBNA4 are the only polypeptides of EBV with separated clusters of unlike charge. This property may be conducive to the formation of a chain of these protein units, each with itself or with other factors.

Charge clusters of a single sign often occur in the first or fourth quartiles (see Table 1), where they are more likely to be on the surface of a globular protein or in an exposed conformation (6). Rawlins *et al.* (10) have suggested that latency is maintained not because the EBNA proteins promote it but because they are negative regulators of the initiation of the lytic cycle. It is tempting to speculate that the EBNA units can form a complex through charge interactions. This aggregation will depend on a high power of the concentrations of the component proteins and may be thus susceptible to an abrupt dissolution of the complex if protein concentration falls.

EBNA1–EBNA3 are known to accumulate in the host cell nucleus. Although none of these proteins has a run of five positive charges shown to be sufficient for this purpose in the case of the large tumor antigen of simian virus 40 (5, 6), it seems possible that their clusters of positive charge may help translocation to the nucleus. In particular, the concentrated positive periodic patterns in EBNA1 and EBNA2 may be especially effective.

(*iii*) *Proteins with multiple charge clusters (positive, negative, or mixed type) and periodic charge patterns.* The conjunction of these properties delimits EBNA1–EBNA3 and the protein product of BMLF1. The gene product of BMLF1 contains two charge clusters, one of negative sign at the amino terminus followed by a mixed charge cluster. The periodic charge patterns in EBNA1 and EBNA2 are embedded in clusters of single charge. Significant runs of the charge pattern $(+, O)_n$ in EBNA1 and EBNA2 emphasize the doublet RG. The glycine molecule is ordinarily not associated with β-sheets or α-helices, but the negligible side chain volume of glycine permits easy movement of its charged neighbors in $(RG)_n$. Therefore, these peptides may adapt conformations to fit well against corresponding charge patterns of opposite sign in another protein.

The periodic charge pattern of BMLF1 is $(R, X, Z)_{10}$, where Z is in eight instances proline and X involves six alanine residues. The predominance of proline in the consensus iteration $(R, A, P)_{10}$ makes it likely that this protein segment may form open coils. This unusual pattern may underlie a novel secondary or tertiary structure and be associated with special regulatory function. The periodic pattern in BMLF1 is followed immediately by RSESRGAGRSTRKQARQERSQRPL involving seven arginine amino acids at displacements of 3 and 4, average 3.43,

Table 1. Significant charge configurations in EBV ORFs

ORF (gene)	Length	f_+	f_-	Cluster					Pattern* or run		
				Location	Length	No. +‖	No. −‖	t	Location		Text
								Latent			
BKRF1 (EBNA1)	641	10.9	7.5	38–75	38	**14**	1	4.61	39–54	$(+, O)_8$	HGRGRGRGRGGGRP
				329–382	54	**21**	3	5.28	354–383	$(+, O)_{15}$	RGRGRERARGGSRERARGRGRGRGEKRPRS
				411–444	34	1	**11**	4.85			
				601–641	41	0	**17**	8.26	621–641	$(O, -, -)_7$	GDDGDDGDEGGDGDEGEEGQE
BYRF1 (EBNA2)	512	10.2	7.4	360–391	32	**14**	1	5.71	369–390	$(O, +)_{11}$	SRGRGRGRGRGRGKGKSRDKQR
				467–501	35	0	**13**	6.71			
BERF1 (EBNA3)	839	11.0	10.5	234–263	30	2	**14**	5.27	739–756	$(-, O, O)_6$	ESGEGSDTSEPCEALDLS
				264–293	30	**11**	0	4.50			
BERF2b (EBNA4)	840	12.6	10.2	50–88	39	**15**	0	4.87			
				247–276	30	1	**17**	7.79	259–263		EDDDE
				304–336	33	**12**	6	4.33	272–276		DEEED
BERF4 (EBNA6)	872	12.6	9.6	243–278	36	1	**13**	4.82			
BNLF1c (LYDMA)	386	7.5	13.2	187–225	39	**11**	**9**	4.71			
				255–299	45	1	**17**	4.43			
								Immediate early lytic			
BMLF1	459	16.3	11.8	6–43	38	1	**17**	5.80	38–43		EEEDED
				89–126	38	**13**	**9**	4.09	130–162	$(+, O, O)_{10}$	RAPRAPRAPRVPRAPRSPRAPRSNRATRGP
BZLF1	200	8.0	6.5	157–199	43	**13**	1	4.81			
								"Early" or "late"			
BPLF1	3149†	11.1	11.1	662–696	35	0	**16**	6.49			
				1395–1444	50	**19**	**12**	6.75			
BLLF1	907	6.6	6.7	193–225	33	1	**9**	4.03			
BRRF2	537	10.4	12.1	463–517	55	3	**24**	5.93			
BKRF4	226	12.4	17.3	53–110	58	1	**31**	6.94			
BBRF3	405	9.4	5.7	351–381	31	**11**	0	4.98	321–332‡	$(+, O, O)_4$	RICRIFKSMRQG
									343–350	$(-, O)_4$	ELELESEP
BXLF1	607	15.2	11.4	20–59	40	2	**15**	4.20			
BVRF1	570	11.4	9.5	40–89	50	**12**	**13**	5.07			
BVRF2	605	12.2	9.6	410–439	30	**7**	**9**	4.18	419–423		EEDEE
BdRF1§	345	11.9	7.8	150–179	30	**7**	**9**	4.63	159–163		EEDEE
BMRF2	357	9.8	3.1	188–217	30	**8**	**4**	4.43			
BFRF1	336	15.5	11.0	251–300	50	**18**	**9**	4.41			

Column 1, the ORF name assigned by Baer *et al.* (2) (biological name is in parentheses); column 2, length of the ORF in residues; column 3, % of positively charged residues; column 4, % of negatively charged residues; column 5, location of cluster in residues; column 6, length of cluster in residues; column 7, number of positive residues in cluster; column 8, number of negative residues in cluster; column 9, significance test (see *Methods*); column 10, location of periodic pattern; column 11, canonical periodic pattern or run of charges; column 12, text of periodic pattern in one-letter code.

*Discrepancies with the pattern are underlined.

†The significance criterion for this long ORF is taken to be 4.5.

‡The significance for $(+, O, O)_4$ is $P \leq 0.016$.

§BdRF1 lies in BVRF2 and is coterminous with it.

‖Numbers in boldface indicate the sign of the cluster.

without intervening prolines, that could readily form an α-helix with a line of positive charge along one side.

BMLF1 is also rich in runs of charged residues near its amino terminus [two negative, $(-)_6$ and $(-)_4$, and two positive, $(+)_5$ and $(+)_4$]. Multiple significant negative runs also occur in EBNA4. The only significant positive-charge run, $(R)_7$, occurs in the gene product of BERF3 [following EBNA4 (BERF2b) on the genomic map].

(*iv*) *Proteins with at least two charge clusters but not of opposite sign and no periodic charge patterns.* LYDMA (BNLF1) and the gene product of BPLF1 fit this description. LYDMA, a plasma membrane protein produced only during latency, contains one negative charge cluster and one of mixed charge. BPLF1 (3149 residues in length) also contains one cluster of negative charge and one of mixed charge. A protein from this very lengthy ORF is unknown. However, it does contribute during latent existence (23, 24) an exon to a lengthy multiexon mRNA ending in EBNA1.

Clusters of charge of any type, because of their tendency to concentrate on the surface of globular proteins, promote the accumulation of less hydrophilic amino acids in the interior of the folded protein. From this perspective it is possible that charge may be as important as hydrophobicity in determining native protein conformation.

(*v*) *Proteins with a single charge cluster.* The immediate early BZLF1 and the late gene BBRF3 possess a single significant cluster of positive charge. BBRF3 is not known to be expressed in latency and is classed late by Baer *et al.* (2). Perhaps it functions like core or capsid proteins that often contain a significant positive cluster or positive run in the fourth (carboxyl) quartile (data not shown). These proteins include L_1 and L_2 genes of human and bovine papilloma viruses, the VP_2 gene of simian virus 40 and polyoma viruses, the pV core protein of adenovirus, and the core gene of human, ground squirrel, and duck hepatitis virus. The positively charged regions in the ends of these capsid proteins may help their accumulation in the nucleus and possibly aid in the organization of the icosahedral capsid about the negatively charged DNA, as may be inferred from the results of Garcea *et al.* (25).

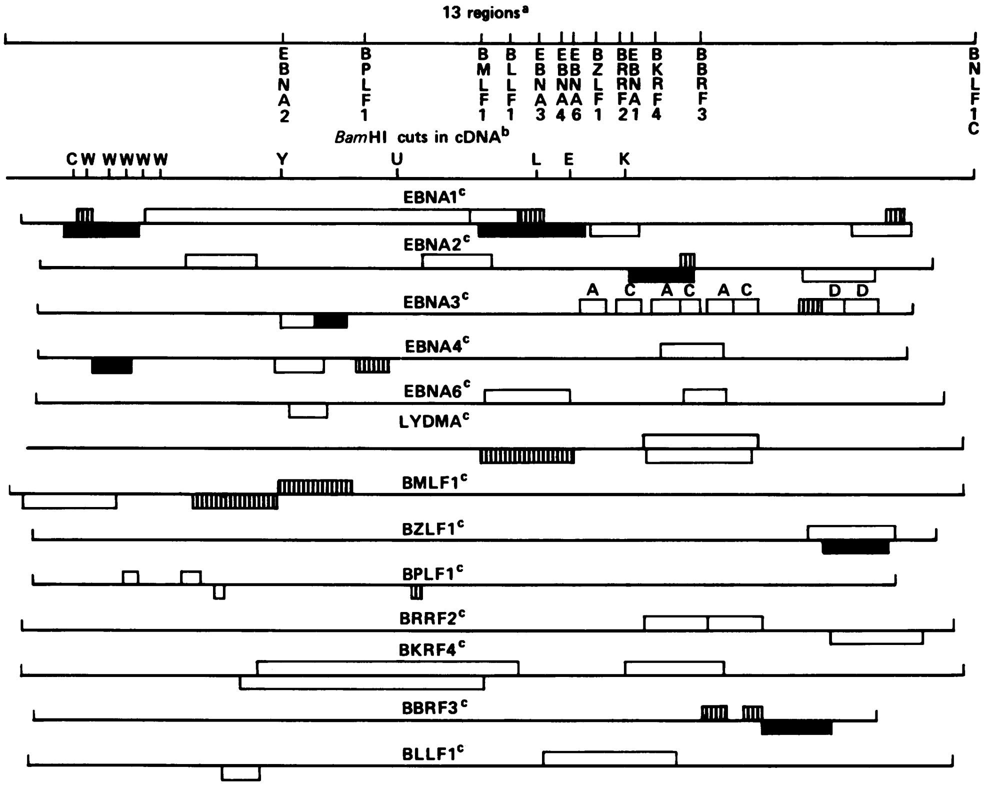

FIG. 1. Relative locations of features in genes having two or more charge or repeat features. Superscript a, locations in 172,282-bp genome of 13 genes having two or more charge or repeat features. Superscript b, locations in the genome of *Bam*HI cuts containing segments (exons) of observed cDNA sequences: W, Y, and EBNA2; W, Y, U, E, and EBNA1; C, W, L, and EBNA3. Superscript c, locations in each of the 13 genes of charge clusters, periodic charge patterns, and repeat regions. Boxes above the line are repeat regions; if barred they are periodic charge patterns. Boxes below the line are charge clusters; if filled they are positive; if clear they are negative; if barred they are of mixed charge.

EBNA6, BLLF1, BRRF2, BKRF4, and BXLF1 each contain a significant single negative charge cluster. BLLF1, like LYDMA, accumulates in the host cell membrane (27), as do many other viral glycoproteins containing clusters of negative charge.

BFRF1, BMRF2, BVRF1, and BVRF2 each contain a single significant mixed charge cluster. BMRF2 has been assigned to a late transcript (2), but no transcript or protein has yet been reported for the others.

(*vi*) ***Proteins with significant periodic charge patterns.*** It is noteworthy that periodic charge patterns only occur in genes with at least one charge cluster of a single sign. The periodic charge patterns are separate from any charge cluster in EBNA3 and BBRF3. The significant periodic charge patterns are quite variable in form. Four are of period 2: $(+, O)_8$ and $(+, O)_{15}$ in EBNA1, $(O, +)_{11}$ in EBNA2, and $(-, O)_4$ in BBRF3. Four are of period 3: $(O, -, -)_7$ in EBNA1, $(-, O, O)_6$ in EBNA3, $(+, O, O)_{10}$ in BMLF1, and $(+, O, O)_4$ in BBRF3. Significant periodic patterns of mixed charge were not observed.

It appears that periodic charge patterns may be associated with some regulatory protein functions. Such a role has been suggested, for example, for the conserved $(+, O, O)_n$, n = 5–8, motif of voltage-gated ion channels; this structure is thought to be a membrane-spanning helix that acts as a voltage sensor and mediates conformational changes in the protein in response to changes in the transmembrane electric field (4). The major immediate early regulatory proteins of the herpes simplex virus (especially ICP-0 and ICP-4) also contain significant periodic charge patterns (data not shown). The major immediate early transcription transactivator protein E1a of adenovirus features the highly significant pattern $(-, O)_6$ = $(EP)_6$ and also some distinctive positive charge patterns.

(*vii*) ***Proteins with significant repeat regions and relationships to charge configurations.*** Fourteen genes have significant charge clusters of a single sign (Table 1), and all except BBRF3 and BXLF1 contain significant repeat elements (mostly tandem copies). Twelve genes in EBV contain substantial oligonucleotide and/or peptide repeats and, notably, all of them have significant charge clusters of a single sign. It is noteworthy that the repeat regions and distinctive charge regions are nonoverlapping except for some in three genes (EBNA1, EBNA2, and BZLF1).

It seems that the conjunction of multiple charge configurations with nonoverlapping repeat regions in the latent and early lytic genes may have a coordinated function. We suggest that these tandem repeat regions may assist in the formation of the multimeric protein complexes bound together by the electrostatic attraction of external charges of unlike sign. Several observations support this proposal. (*a*) Except for fibrous, occasional structural, and some regulatory proteins, repeats are rare in proteins (28). (*b*) The repeats in EBNA1–EBNA3 are variable in length in different virus strains (29), which makes it unlikely that they lie in compact well-structured globular domains. (*c*) The protein size estimated from mobility in gels relative to globular protein standards is considerably larger than that inferred from the ORF text (30), possibly implying that these proteins adopt an unusually extended conformation. (*d*) The long repeat in EBNA1 is very high in glycine and those in EBNA2–EBNA4 and EBNA6 are very high in proline, both amino acids prominent in open coil structures and lacking in α-helices or β-strands (31). (*e*) Many of the long peptide repeats have few substitutions at the DNA level, implying that they may be of recent amplification and that their detailed structure may not be of great functional importance. These extended repeat regions might act as flexible open coil domains linking the functional globular domains. Similar poorly conserved regions rich in proline lying between two distinct highly conserved functional domains have been proposed to be a flexible hinge in the mineralocorticoid receptor and in other sequentially similar molecules (32). They would make it easier for the protein to adopt conformations favorable to interaction with charged regions on other proteins in the formation of multimeric protein complexes. This flexibility could be particularly helpful in EBNA1, where the long (glycine, alanine) repeat lies between charge cluster regions.

(*viii*) *Miscellany*. In view of the prominent charge configurations in EBV proteins associated with the maintenance or disruption of the latent state, it is interesting to compare them to the proteins of varicella-zoster virus (VZV), a herpesvirus that is not known to produce proteins in the latent state (after primary infection, VZV lives latently in generally nonreplicating nerve cells). The complete sequence of a strain of VZV has been determined (33). Only 6 of its 68 distinct ORFs contain significant positive or negative charge clusters, and only one, gene 11, contains charge clusters of each sign. Gene 11 also has a significant repeat region, coincident with the negative charge cluster. The other proteins of VZV with prominent repeats do not present distinctive charge configurations. The relative absence in VZV of multiple charge configurations and repeats, and their prominence in proteins of EBV associated with the latent state or its disruption, reinforces the conclusion that these features are of functional importance to the maintenance and termination of the latent state. A detailed analysis of the charge distribution in other viruses of the herpes family (34) and for other classes of eukaryotic and prokaryotic protein sequences will be presented elsewhere.

It is interesting to speculate about the mechanisms that bind to DNA the *n*-mers of the carboxyl-terminal 191-residue EBNA1 fragment studied by Milman and Hwang (22). We suggest that the segment $(PG)_4P$ occurring in the middle of the 191-residue sequence is very likely to promote an open coil region that can protrude from the surface of the *n*-mer and be bound to the DNA as does an exposed loop of DNase (35). Protrusion of a loop may be promoted by ionic binding of a short concentration of positive residues at the 5′ end of the protein to the negative charge cluster (17 negative residues in 41; no positives) at the 3′ end of the same protein. Since ionic forces are long range, the repulsion of the excess negative charges at the 3′ end by the negative phosphate exterior of the DNA may aid in proper orientation of the loop for binding to the DNA. Of course, attraction between the unlike charges on different molecules can form chains of *n* units.

The many striking significant charge configurations we found in these few proteins and the possible functions suggested for them invite experimental investigation. For example, is the $(RXY)_{10}$ periodic pattern (X mostly alanine and Y involving proline eight times) in BMLF1 vital to its transactivation function for the lytic cycle? How essential are its two charge clusters? Such questions can be investigated by attenuation, deletion, or relocation and other sequence manipulations. The importance of the repeat charge pattern can be investigated by reconstructions such as replacing $(RAP)_{10}$, where A promotes α-helices, P promotes open coils, and R is positively charged, by $(KAP)_{10}$ (a different positive more favorable to open coil), by substituting E for R (negative) or T for R (neutral), where E and T have about the same open coil-promoting propensity as R, or by $(RAA)_{10}$, which would promote an α-helix structure, or $(RLL)_{10}$, which could lead to an amphipathic helix structure.

We are happy to acknowledge valuable comments on this manuscript from Drs. E. Mocarski, M. Calos, and R. Jariwalla and from Macdonald Morris. This work was supported in part by National Institutes of Health Grants GM10452-22 and 5HL30856-02, National Science Foundation Grant MCS82-15131 to S.K., and National Institutes of Health Grant R01GM3056-01 to B.E.B.

1. Miller, G. (1985) in *Virology*, eds. Fields, B. N., Knipe, D. M., Melnick, J., Chanock, R. M., Roizman, B. & Shope, S. E. (Raven, New York), pp. 536–589.
2. Baer, R., Bankier, A. T., Biggin, M. D., Deiniger, P. L., Farrel, P. J., Gibson, T. J., Hatfull, G., Hudson, G. S., Satchwell, S. C., Sequin, C., Tuffnell, P. S. & Barrell, B. G. (1984) *Nature (London)* **310**, 207–211.
3. Karlin, S. & Blaisdell, B. E. (1987) *J. Mol. Evol.* **25**, 215–229.
4. Noda, M., Shimizu, S., Tanabe, T., Takai, T., Kayano, T., Ikeda, T., Takahashi, H., Nakayama, H., Kanaoka, Y., Minamino, N., Kangawa, K., Matsuo, H., Raftery, M. A., Hirose, T., Inayama, S., Hayashida, H., Miyata, T. & Numa, S. (1984) *Nature (London)* **312**, 121–127.
5. Richardson, W. D., Roberts, B. L. & Smith, A. E. (1986) *Cell* **44**, 77–85.
6. Smith, A. E., Kalderon, D., Roberts, B. L., Colledge, W. H., Edge, M., Gillett, P., Markham, A., Paucha, E. & Richardson, W. D. (1985) *Proc. R. Soc. Lond. Ser. B* **226**, 43–58.
7. Spangler, R., Bruner, M., Dalie, B. & Harter, M. L. (1987) *Science* **237**, 1044–1048.
8. Karlin, S. & Ost, F. (1987) *Adv. Appl. Probab.* **19**, 293–351.
9. Reisman, D., Yates, J. & Sugden, B. (1985) *Mol. Cell. Biol.* **5**, 1822–1832.
10. Rawlins, D. R., Milman, G., Hayward, S. D. & Hayward, G. S. (1985) *Cell* **42**, 859–868.
11. Reisman, D. & Sugden, B. (1986) *Mol. Cell. Biol.* **6**, 3838–3846.
12. Skare, J., Farley, J., Strominger, J. L., Fresen, K. O., Cho, M. S. & zurHausen, H. (1985) *J. Virol.* **55**, 286–297.
13. Bodescot, M., Brison, O. & Perricaudet, M. (1986) *Nucleic Acids Res.* **14**, 2611–2620.
14. Bodescot, M., Perricaudet, M. & Farrell, P. J. (1987) *J. Virol.* **61**, 3424–3430.
15. Ricksten, A., Kallin, B., Alexander, H., Dillner, J., Fahraeus, R., Klein, G., Lerner, R. & Rymo, L. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 995–999.
16. Dillner, J., Kallin, B., Alexander, H., Ernberg, I., Uno, M., Ono, Y., Klein, G. & Lerner, R. A. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 6641–6645.
17. Fennewald, S., van Santen, V. & Kieff, E. (1984) *J. Virol.* **51**, 411–419.
18. Chevallier-Greco, A., Manet, E., Chavrier, P., Mosnier, C., Daillie, J. & Sergeant, A. (1986) *EMBO J.* **5**, 3243–3249.
19. Countryman, J., Jenson, H., Seible, R., Wolf, H. & Miller, G. (1987) *J. Virol.* **61**, 3677–3679.
20. Lieberman, P. M., O'Hare, P., Hayward, G. S. & Hayward, S. D. (1986) *J. Virol.* **60**, 140–148.
21. Manet, E., Chevallier, A., Zhang, C. X., Ooka, T., Chavrier, P. & Daillie, J. (1985) *J. Virol.* **54**, 608–614.
22. Milman, G. & Hwang, E. S. (1987) *J. Virol.* **61**, 465–471.
23. Sample, J., Hummel, M., Braun, D., Birkenbach, M. & Kieff, E. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 5096–5100.
24. Speck, S. H. & Strominger, J. L. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 8305–8309.
25. Garcea, R. L., Salunke, D. M. & Caspar, D. L. D. (1987) *Nature (London)* **329**, 86–87.
26. Davison, A. J. & Taylor, P. (1987) *J. Gen. Virol.* **68**, 1067–1079.
27. Beisel, C., Tanner, J., Matsuo, T., Thorley-Lawson, D., Kezdy, F. & Kieff, E. (1985) *J. Virol.* **54**, 665–674.
28. Dayhoff, M. O. (1979) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3.
29. Hennessy, K., Wang, F., Bushman, E. W. & Kieff, E. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 5693–5697.
30. Hennessy, K. & Kieff, E. (1985) *Science* **227**, 1238–1240.
31. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13**, 222–244.
32. Arriza, J. L., Weinberger, C., Cerelli, G., Glaser, T. M., Handelin, B. L., Housman, D. E. & Evans, R. M. (1987) *Science* **237**, 268–275.
33. Davison, A. J. & Scott, J. E. (1986) *J. Gen. Virol.* **67**, 1759–1816.
34. Karlin S., Blaisdell, B. E. & Brendel, V. (1988) *J. Mol. Biol.*, in press.
35. Suck, D., Lahm, A. & Oefner, C. (1988) *Nature (London)* **332**, 464–468.