Charge Configurations in Viral Proteins

Samuel Karlin; Volker Brendel

Proceedings of the National Academy of Sciences of the United States of America, Volume 85, Issue 24 (Dec. 15, 1988), 9396-9400.

Stable URL:
http://links.jstor.org/sici?sici=0027-8424%2819881215%2985%3A24%3C9396%3ACCIVP%3E2.0.CO%3B2-7

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 9396–9400, December 1988
Biochemistry

Charge configurations in viral proteins

(charge clusters/charge patterns/regulatory proteins)

SAMUEL KARLIN* AND VOLKER BRENDEL

Department of Mathematics, Stanford University, Stanford, CA 94305

Contributed by Samuel Karlin, August 25, 1988

ABSTRACT **The spatial distribution of the charged residues of a protein is of interest with respect to potential electrostatic interactions. We have examined the proteins of a large number of representative eukaryotic and prokaryotic viruses for the occurrence of significant clusters, runs, and periodic patterns of charge. Clusters and runs of positive charge are prominent in many capsid and core proteins, whereas surface (glyco)proteins frequently contain a negative charge cluster. Significant charge configurations are abundant in regulatory proteins implicated in transcriptional transactivation and cellular transformation. Proteins with charge structures are much more predominant in animal DNA viruses as compared to animal RNA viruses and prokaryotic viruses. This contrast might reflect the role of protein charge structures in facilitating competitive virus–host interactions involving the cellular transcription, translation, protein sorting, and transport apparatus.**

Ionic interactions are of importance in relation to protein structure, function, and cellular processes. Highly charged segments and periodic arrangements of charged residues have been implicated in DNA binding and transcriptional activation (1, 2), in transmembrane ion transport (3, 4), and in nuclear location signals and protein sorting (5, 6). We have recently introduced methods to characterize distinctive charge configurations in the primary structure of a protein and to assess their statistical significance given the amino acid composition (7). Our methods identify long uninterrupted sequences ("runs") of charge, charge "clusters" (about 30- to 50-residue segments with unusually high specific charge content), and various periodic patterns of charge, such as $(-,0)_n$, $(+,0,0)_n$, $(+,0,-,0)_n$, and $(0,-,-)_n$, where $+$, $-$, and 0 designate a positively, negatively, and uncharged amino acid, respectively. Charge patterns of period 2 occurring in a β-sheet would present a straight line of charge on one side of the sheet, and patterns of period 3–4 would approximately align the charges on one side of an α-helix. The occurrence of one or more charge clusters or patterns within a protein could contribute to establishing chains or complexes of such protein units as well as to cooperative protein–protein and protein–nucleic acid interactions or might serve in intramolecular folding.

Application of our methods to the known and putative proteins of human herpesviruses [Epstein–Barr virus (EBV), varicella–zoster virus (VZV), herpes simplex virus (HSV), and cytomegalovirus (CMV)] revealed a richness of charge configurations associated with certain groupings of these proteins (7). In particular, many membrane-associated proteins of these viruses exhibit a clustering of negatively charged residues, and the immediate early gene products active in transcriptional regulation commonly contain multiple distinctive charge configurations. In EBV, the products of 14 of the 84 identified open reading frames (ORFs) carry at least one significant charge cluster. Twelve ORFs of EBV involve significant repeats and also contain significant charge clusters of a single sign. The repeat regions and charge clusters are mostly separate, and complementary functions of these features have been suggested (8). Eight of these 12 ORFs produce proteins serving during latency or initiating the lytic cycle. The latency-associated major nuclear proteins of the latent state (EBV-encoded nuclear antigens 1, 2, 3, and 4) are the only EBV proteins containing separate charge clusters of opposite sign. The exclusive occurrence of multiple charge configurations and repeats in proteins of EBV associated with the latent state or its disruption supports the hypothesis that these features are of functional importance in the maintenance and curtailment of the latent state (8).

In this paper we continue with the analysis of the charge distribution for a broad spectrum of protein sequences derived from animal, plant, and bacterial viruses. The results reveal a number of common themes as well as distinct contrasts in the nature and extent of charge configurations occurring in proteins of different functions and origins. Particular charge structures tend to be associated with capsid-core proteins, surface proteins, and regulatory proteins. Polypeptides with charge structures predominate in mammalian DNA viruses compared to RNA viruses, prokaryotic viruses, and sets of host proteins.

RESULTS

We previously reported statistically significant charge configurations for the viral proteins of the herpesvirus family (7). Distinctive charge patterns were found in the products of about 20% of the ORFs in EBV, 10% in VZV, 35% in HSV1, and ≈25% of those available in CMV. Tables 1–4 display the significant charge configurations found in the known and putative protein sequences in >35 other DNA and RNA viral genomes, including representative human and animal viruses, plant viruses, and several prokaryotic phages. Significance was estimated according to methods explained in detail in refs. 7 and 9; all charge configurations reported here are statistically significant at the 1% level.

Animal DNA Viruses. The animal DNA viruses as a group have many proteins with significant charge configurations.

Adenoviruses. Adenoviruses are nonenveloped icosahedral virions containing a linear double-stranded DNA molecule, 36–40 kilobase pairs (kbp). For adenovirus type 2 the sequences of 28 putative and known proteins have been reported. Table 1 displays all significant charge configurations found in these protein sequences.

The early gene products E1–E4 are active soon after infection; E1 expression precedes that of E2–E4 and plays a

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; VZV, varicella–zoster virus; HSV, herpes simplex virus; CMV, cytomegalovirus; ORF, open reading frame; SV40, simian virus 40; T and t, large tumor and small tumor; CaMV, cauliflower mosaic virus; Mo-MLV, Moloney murine leukemia virus; HIV, human immunodeficiency virus; RB, retinoblastoma susceptibility gene product.

*To whom reprint requests should be addressed.

Table 1. Significant charge configurations in adenovirus proteins

Protein	Charge configuration
E1A 32K	$(-)_9$ DDEDEEGEE 133–141
(289 11.1 17.3)	$(+, 0, 0)_5$ HPGHGCRSCHYHRRN 149–163
	$(-, 0)_6$ EPEPEPEPEPEP 189–200
DNA polymerase	$(+, 0)_5$ KGKLRAKGHA 953–962
(1056 14.7 13.2)	
Terminal protein	$(+)_6$ RRRRRR 366–371
(653 14.1 13.5)	− cluster (2, 17/33) 380–412
52–55K	± cluster (13, 14/42) 100–141
(415 16.6 16.4)	
IIIa	± cluster (9, 9/33) 529–561
(585 12.0 11.6)	
pro-VII	$(+)_{14}$ KRRRRRVARRHRRR 99–112
(198 26.3 4.0)	
pV	± cluster (12, 12/30) 23–52
(369 20.9 17.5)	+ cluster (15, 0/28) 310–339
Hexon	$(-)_{16}$ EDEEEEDEDEEEEEEE 147–162
(968 10.3 11.8)	
DBP	$(+)_6$ KKKKKR 85–90
(529 16.6 12.9)	
100K	− cluster (3, 19/40) 39–78
(805 13.5 13.9)	
33K	− cluster (0, 22/39) 16–54
(228 14.0 14.9)	

The number of residues and the proportions of positively and negatively charged residues are given below the name of each protein. Significant charge configurations are classified by type and displayed in the standard one-letter code. For each cluster the number of basic and acidic residues and the length of the cluster are given in that order. The residue coordinates manifesting the charge configuration are indicated. DBP, DNA binding protein.

regulatory role in the expression of the latter. Through differential splicing, E1A comprises at least five mRNAs encoding several proteins, of which 32K (289 residues) and 26K (243 residues), required for efficient expression of all the other adenovirus genes, are the largest. The E1A proteins, which have transactivating and transforming functions (for review, see ref. 10), are rich in charge structures. The sequences of E1A 26K and E1A 32K share the charge run DDEDEEGEE (the underlined residue indicating a mismatch from the pattern) and the unusual glutamate-proline iteration $(EP)_6$; see Table 1. The run of negatively charged residues is highly conserved among E1A proteins of different adenovirus strains (11). The E1A protein segment residues 121–139, which contains the aforementioned acidic charge run, is assigned an important role in transformation (12, 13).

The iteration $(EP)_6$ residues 189–200, in view of the proline component, presents a peptide not conducive to any regular secondary structure. Attesting to its possible functional importance is the fact that the EP iteration is not simply due to a DNA repeat, the proline residues being translated from all four proline codons and the glutamate residues deriving from both glutamate codons. The two acidic sequences in E1A 26K are essentially juxtaposed, but in E1A 32K these sequences are separated by the period 3 positive charge run $(+, *, *)_5$ = HPGHGCRSCHYHRRN emphasizing histidine and including two cysteine components. The contiguous sequence of the negative charge run, the concentrated segment of positive charge, and the reiterations of the glutamate-proline doublet constitute the major effector domain responsible for E1A-inducible transcriptional control (10, 12, 13). The 105K cellular protein found in complex with E1A was identified as the retinoblastoma susceptibility gene product (RB); thus, E1A-mediated immortalization might be due to inactivation of RB, a putative neoplastic inhibitor (14). Parenthetically, RB also carries a significant acidic charge cluster in its N quartile.

The role played by E1B protein in transformation is unclear. E1A at high levels can transform cells in the absence of E1B (10). Although the E1A proteins interact with E1B, no statistically significant charge configurations are discerned for E1B. At its C terminus, E1B features the short charge segment EEQQQEEARRRRRQEQ.

Several adenovirus proteins involved in DNA replication have significant charge configurations (see ref. 15 for a review on adenovirus). The terminal protein (TP) of adenovirus is covalently attached to the 5′ ends of the viral DNA and is essential for initiation of replication (15). In its third quartile the protein displays a run of six arginines followed by a negative charge cluster comprised of considerable local iterations RRRRRRVPPPPPPPEEEEEGEALMEEEIEEEE (residues 366–397). The 72-residue range 349–420 of TP is highly charged, which conceivably helps order and stabilize the assembly of the adenovirus in the nucleus. The adenovirus DNA binding protein, an early product necessary for the elongation of adenovirus DNA (15), contains a 6-mer of positively charged residues (see Table 1). The DNA polymerase carries the statistically significant alternating pattern $(+, 0)_5$ = KGKLRAKGHA.

The 52–55K protein features a mixed charge cluster in its second quartile (see Table 1) and the negative charge peptide EEYDEDDEYEPEDGEY at the C terminus; the function of the 52–55K polypeptides is unknown.

Protein pro-VII is the 198-residue precursor to the major core protein VII. Besides the long protamine-like internal run of positive residues, pVII also carries a large overall net positive charge (see Table 1), consistent with a histone-like function. A structure has been suggested for pVII wherein four hydrophobic α-helices (two entailing long iterations of alanines) separate basic domains; these basic domains are thought to interact with DNA phosphates, whereas hydrophobic interactions between the α-helical segments are thought to affect condensation of the viral DNA analogous to nucleoprotamine (16).

Protein V binds to linker-like segments of the viral DNA and can also bind to a penton base; it might provide a bridge between virion DNA and the virion capsid (15). pV has a mixed charge cluster near its N terminus, including the charge run $RKLKRVK_5D_5ELD_2EVE$ (residues 32–54), which exhibits a sharp switch of positive- to negative-charged residues. Near its C terminus, pV contains an R_7 run within an extended segment of highly basic character (see Table 1), typical of core proteins of other double-stranded DNA viruses (see sections on hepatitis B virus and papovaviruses below).

The adenovirus hexon protein contains a remarkable run of 16 acidic residues (see Table 1). It is interesting that also the adenovirus 33K nonstructural protein of unknown function and the 100K protein, which is thought to provide a scaffold for hexon trimer assembly, carry negative charge clusters towards their N termini, whereas the hexon-associated protein IIIa carries a mixed charge cluster at its C terminus. One might hypothesize that this common pattern is recognized by some positively charged factor involved in virion assembly, possibly even the core proteins V and VII.

Parvoviridae. Of this group of single-stranded DNA viruses, the genomes of minute virus of mice, an autonomous parvovirus, and of adeno-associated virus 2 (AAV2), a virus dependent for replication on the coinfection of either adenovirus or herpesvirus, have been completely sequenced. The left-half ORF of AAV2 encodes *rep* proteins involved in regulation of gene expression and inhibition of viral replication (17) and contains the only significant charge structure in the set of parvovirus polypeptides (see Table 2).

Hepatitis B virus. The genome of human hepatitis B virus encodes the core antigen, a viral DNA polymerase/reverse transcriptase, and a 389-residue precursor that is processed

Table 2. Significant charge configurations in proteins of animal DNA viruses

Virus	Charge configuration
Adeno-associated 2 (AA2 4.7 kbp 4)	
Left-hand ORF: 621 14.0 12.4	± cluster (20, 14/68) 443–510
Human hepatitis B (HPBAYW 3.2 kbp 4)	
Core antigen: 183 16.4 9.8	+ cluster (16, 0/30) 150–179
Surface antigen: 389 5.7 3.3	± cluster (4, 6/31) 16–46
DNA polymerase: 832 15.5 5.6	− cluster (3, 10/30) 11–40
SV40 (SV4CG 5.2 kbp 7)	
VP2 (VP3): 352 10.8 9.4	+ cluster (16, 0/33) 319–351
Large-T antigen: 708 15.4 14.8	$(-)_6$ DDDDED 633–638
Mouse polyoma (PLY2 5.3 kbp 6)	
Middle-T antigen: 421 15.4 11.4	− cluster (0, 14/31) 294–324
Human papilloma 33 (PPH33CG 7.9 kbp 8)	
E1: 648 11.9 14.0	− cluster (0, 14/31) 38–68
E7: 107 9.3 13.1	− cluster (0, 11/30) 20–49
L1: 525 12.4 9.7	+ cluster (14, 0/35) 491–525
L2: 471 10.4 9.1	+ cluster (13, 0/36) 292–327 $(+)_6$ RRRRKR 454–459
Vaccinia (185 kbp 50)	
pA5L: 281 8.9 11.0	± cluster (7, 10/35) 25–59
pC1L: 244 11.9 19.3	$(-, 0)_9$ EITESESDPDPEVE-SEDDS 52–70
pJ3R: 333 16.8 9.9	$(-, 0)_8$ EIDNELDYEPESAN-EV 15–30

The GenBank identification, size, and number of ORFs are given for each virus. Displayed for each protein are its number of residues, the proportions of positively and negatively charged residues, the type of charge configuration, the coordinates of the charge configuration, and either the sequences in the one-letter code (mismatches underlined) or (for clusters) the number of basic and acidic residues and the length of the cluster, in that order.

to several surface antigens (S); the product of another 154-residue ORF (X) has not been identified. The core antigen features a charge cluster of basic residues proximal to the C terminus, the pre-S product contains a mixed charge cluster near its N terminus, and the polymerase has a negative charge cluster at its N terminus (see Table 2). These features are conserved in the corresponding proteins of the ground squirrel hepatitis B virus and of woodchuck hepatitis B virus. The duck hepatitis B virus core protein, although much less similar to the corresponding human and ground squirrel core proteins, also terminates in a high basic charge concentration (data not shown).

Polyoma, Simian virus 40 (SV40). Papovaviruses, which carry a double-stranded circular DNA genome of about 5.2 kbp contained in an icosahedral capsid, are capable of cell immortalization and transformation. SV40, a well-studied member of this family, encodes seven polypeptides, including the early transcribed large-tumor (T) and small-tumor (t) antigens and three late proteins, VP1, VP2, and VP3. Mouse polyomavirus is similarly arranged but also expresses a middle-T antigen. SV40 large-T antigen is found mostly in the nucleus; it is autoregulative, stimulates cellular DNA synthesis, is involved in initiating viral DNA synthesis, and is required for late gene expression. Polyoma large-T antigen can induce indefinite growth in primary cells; middle-T antigen alone can transform rat 3T3 cell lines.

Charge analysis of the SV40 polypeptides reveals a pronounced positive charge cluster at the C terminus of VP2 (see Table 2). The N terminus of VP2 is very hydrophobic, containing a significantly long run of 43 noncharged residues. VP2 and VP3 are capsid components of the SV40 virion known to be associated with the viral DNA, possibly by way of the basic C terminus (18).

The SV40 large-T antigen displays a short run of positively charged residues PKKKRKV (positions 126–132), which has been associated with a signaling function for nuclear location (5). Near its C terminus, the T antigen also contains a hexamer of acidic residues (see Table 2), contained in an almost significant acidic charge cluster (0.02 probability level). Large-T antigen is found in complex with p53, AP_2, DNA polymerase-α, and RB; these associations are thought to be instrumental in T-antigen-induced transformation (19). It is intriguing that in these complexes the component proteins carry at least one distinctive charge configuration.

The positive charge cluster in VP2 is preserved in the same relative location in the homologous proteins of the human papovaviruses as is the run of uncharged residues. Mouse polyomavirus VP2 has a run of five basic residues at the C terminus but no significant charge cluster. The middle-T antigen of polyoma contains a strong cluster of negatively charged residues (see Table 2).

Papillomaviridae. Four of the seven substantial ORFs of human papillomavirus type 33 translate into proteins with significant charge configurations (see Table 2). Protein E1, which contains a negative charge cluster, is involved in DNA replication control. Also, E7 carries a strong negative charge cluster. HPV-16 E7 has been shown to have transactivating and transforming functions similar to those of adenovirus E1A (20). The products of the late genes L1 and L2 are capsid proteins, both featuring positive charge clusters at the C terminus. These charge patterns are mostly conserved between different virus strains (data not shown).

Poxviruses. Vaccinia virus, a representative of the family of poxviruses, has a genome of about 185 kbp. Vaccinia encodes enzymes for transcription, capping, polyadenylylation, and replication. Viral replication occurs in the cytoplasm although not in an enucleated cell. Nuclear enzymes are apparently not required for viral replication or mRNA production. Virus assembly also occurs in the cytoplasm.

The sequences of some 50 putative vaccinia polypeptides are available (B. Moss, personal communication). The gene product of A5L, possibly a subunit of the viral RNA polymerase, contains a mixed charge cluster close to the N terminus (see Table 2) as well as a 75-residue stretch of uncharged residues toward the C terminus. The gene product of C1L (a late gene) contains a significant $(-, 0)_9$ periodic charge pattern where the 0 components are predominantly serine or proline, and a similar pattern is found in pJ3R (see Table 2).

Animal RNA Viruses. The animal RNA viruses as a group have few proteins with significant charge configurations.

Picornaviridae. This group of small isometric virions includes poliovirus type 1, human hepatitis A virus, and human rhinovirus type 14. Among the mature proteins of these viruses only the VP1 + VP4 capsid protein of human hepatitis A virus features a significant charge configuration (a mixed charge cluster; see Table 3).

Togaviridae. Three representative viruses of this group have been completely sequenced: Sindbis virus, yellow fever virus, and West Nile virus. No statistically significant charge configurations occur in any of their gene products; the nonstructural protein P3 in the Sindbis virus, however, contains the substantial positive run RKQRRRRRSRR at the carboxyl end, reminiscent of corresponding charge runs of other viral capsid proteins.

Retroviridae. Five representative retrovirus genomes were examined (see Table 3): Rous sarcoma virus, Moloney murine leukemia virus (Mo-MLV), human immunodeficiency virus (HIV; isolate HXB2), visna lentivirus, and human T-cell leukemia virus type 1. An extraordinary mixed charge cluster occurs at the C terminus of the p30 protein of Mo-MLV, involving 33 charged residues in a segment of length 36. p30 is a nucleocapsid protein that is thought to be

Table 3. Significant charge configurations in proteins of animal RNA viruses

Virus	Charge configuration
Human hepatitis A	
(HPAACG 7.5 kbp 10)	
VP1 + VP4: 345 12.5 12.2	± cluster (11, 9/34) 289–322
Mo-MLV	
(MLM 5.8 kbp 9)	
p30: 263 18.3 17.5	± cluster (16, 16/36) 223–258
HIV	
(isolate HXB2 9.7 kbp 7)	
gag polyprotein: 500 15.4 10.8	± cluster (12, 5/32) 12–43 (+, X, X)$_{13}$ KIVKCFNCGKEGHTARNCRAPRKKGCWKCGKEGHQMKDC 388–426
Vesicular stomatitis	
(VSV 11.2 kbp 5)	
NS: 265 11.7 18.9	(−, 0)$_6$ DSDTESEPEIEDN 59–71
Influenza A (13.6 kbp 10)	
Polymerase 1: 757 15.1 10.6	+ cluster (15, 1/32) 184–215

Entries are as explained in the legend to Table 2. No significant charge configurations occur in the proteins of poliovirus type 1, human rhinovirus type 14, Sindbis virus, yellow fever virus, West Nile virus, Rous sarcoma virus, visna lentivirus, human T-cell leukemia virus type 1, and reovirus type 1.

the essential protein in the gag complex since mutations in p30, but not in the other gag genes, are lethal to the virus (21).

The gag polyprotein of HIV features two almost significant mixed charge clusters (with 2% statistical confidence) in close proximity near the N terminus: residues 12–43, containing 12 basic and 5 acidic residues, and residues 89–118, containing 9 basic and 7 acidic residues. The gag polyprotein also has a long period-3 charge pattern with two mismatches (+, X, X)$_{13}$ (see Table 3) at coordinates 388–426. The HIV polymerase contains the significant period-3 positive charge pattern (+, 0, 0)$_6$ = KLNKTGKYARMRGAHTN with one deletion. The C terminus of the tat protein, considered to function as transactivator in the nucleus, presents the substantial (though not statistically significant) positive charge run RKKRRQRRR.

Rhabdo-, orthomyxo-, and reoviridae. The only gene product of this group displaying any significant charge clusters is polymerase 1 of influenza virus A, which has a positive charge cluster in the first quartile (see Table 3). The surface protein hemagglutinin has a significant 39-residue uncharged run over the coordinate range 524–562 proximal to the C terminus. The NS protein of vesicular stomatitis virus contains a (−, 0)$_6$ pattern (see Table 3).

Plant Viruses. We have examined the proteins of the DNA viruses maize streak virus, wheat dwarf virus, and cauliflower mosaic virus (CaMV) and the proteins of the RNA viruses tobacco mosaic virus, tobacco etch virus, satellite tobacco necrosis virus, and alfalfa mosaic virus. The only virus encoding a protein with a charge configuration of note is CaMV. CaMV is an icosahedral virion that is generally propagated by aphids. Its genome replication cycle involves reverse transcription of an mRNA template to yield a RNA·DNA hybrid that is subsequently converted to a double-stranded DNA identical to the parental genome. All CaMV proteins are rich in lysine, and proteins III and VI are also high in proline (22). The coat protein CaMV (length, 488 residues: 17.2% basic, 18.9% acidic) displays a pronounced negative charge cluster (coordinates 449–488: 1 basic and 19 acidic residues in 40) preceded by a pronounced positive charge cluster at the C terminus (coordinates 357–408: 30 basic and 1 acidic residues in 52). The positive charge cluster contains the sequence KKTS KKK Y HKR Y KKR Y KV Y KP Y KKKKK with striking predominance of tyrosine separating short runs of mainly lysines. One might speculate the tyrosines to be phosphorylation sites with the lysines providing an ionic milieu suitable for kinase activity. Proteins like the CaMV coat protein, which display separated clusters of opposite charge, might possibly assemble into multimers due to electrostatic attraction.

Table 4. Significant charge configurations in proteins of bacteriophages

Phage	Charge configuration
λ (LAM 48.5 kbp ≈50)	
Tail Z (192 25.0 4.7)	(+)$_7$ RRRRRKK 94–100
Tail V (246 8.9 10.6)	− cluster (0, 10/30) 44–73
Tail H (853 13.2 11.8)	± cluster (13, 12/50) 326–375
Tail J (1132 11.4 11.4)	± cluster (13, 17/67) 809–875
Nin204 (204 25.0 13.2)	± cluster (17, 8/33) 40–72
T7 (PT7 39.9 kbp ≈60)	
Kinase 0.7 (359 17.5 15.9)	± cluster (24, 11/52) 231–282
PZA (PZACG 19.4 kbp 23)	
Tail p (599 11.7 11.0)	± cluster (14, 1/37) 515–551

Entries are as explained in the legend to Table 2. No significant charge patterns occur in the proteins of phages φX174, f1, and MS2.

Bacterial Viruses. The charge configurations of polypeptides of phages φX174 (11 proteins), f1 (9 proteins), λ (about 50 proteins), T7 (about 60 proteins), PZA (23 proteins), and MS2 (4 proteins) (all of *Escherichia coli* except PZA of *Bacillus subtilis*) were investigated. Of the 60 known or potential proteins of T7, only the kinase protein contains a charge cluster. For the λ proteins, the proteins with significant charge structures are all tail components (Table 4; the function of Nin204 is unknown). In all other phage sequences examined we found but a single charge cluster near the C terminus of a tail protein from PZA. The charge patterns in these tail proteins are not of an invariant form.

DISCUSSION

Several intriguing contrasts and observations on the nature and extent of significant charge configurations stand out with respect to proteins of DNA versus RNA eukaryotic viruses, of eukaryotic viral proteins versus prokaryotic viral proteins, and among classes of proteins. On average about 20% of the proteins in double-stranded DNA mammalian viruses, independent of genome size, compared to less than on average 3% of RNA viral proteins, contain some significant charge sequences. The paucity of charge structures in bacteriophage proteins compared with the abundance of charge structures in the proteins of eukaryotic DNA viruses is also striking. What about protein function and structure in relation to charge configurations? Capsid, core, and coat proteins in DNA viral genomes often feature a positive charge cluster or run generally near the C terminus, many membrane-associated and glycoproteins contain a negative charge cluster, and transactivating proteins tend to have multiple significant charge configurations.

The capsid protein VP2 of SV40, the major and minor capsid proteins of human papillomavirus, the core protein pV of adenovirus, and the core protein of the hepatitis B virus all carry significant clusters or runs of positive charge at or near their C terminus. Possible roles for these charge structures might include nuclear accumulation, association with viral DNA, and orientation of the capsid proteins one to the other during assembly of the virion coat (23).

The rare cases of charge clusters in RNA viruses occur principally in core-capsid proteins. Thus, the VP1–VP4 capsid protein of hepatitis A virus and the p30 nucleocore of Mo-MLV contain mixed charge clusters, and the gag (core-like) protein of HIV contains two distinct mixed charge clusters, all near the C or N terminus. The coat protein of CaMV has two pronounced proximal charge clusters of

opposite sign near the C terminus. These charge clusters of opposite sign may be conducive to the formation of a chain of these protein units with each other or with other factors stabilized through charge–charge interactions.

Many viral glycoproteins and surface proteins feature a negative charge cluster. Such is the case for gD, gE, and gG of HSV, for LMP and gp350 of EBV, for protein 68 of VZV (7), and for the adenovirus hexon protein. The negative charge cluster might facilitate association to the host cell wall.

Transcription activators and transforming proteins tend to be associated with multiple charge structures, often including mixed charge clusters. This holds true for E1A of adenovirus, large-T antigen of SV40 and middle-T antigen of polyoma, E7 of papillomavirus, and the principal immediate early transactivators of the human herpesviruses [pBMLF1 of EBV, p62 and p63 of VZV, infected cell proteins 0 and 4 (ICP-0 and ICP-4) of HSV-1, and IE1 of CMV]. The same motif occurs in many cellular transcriptional activators, including yeast GCN4, PHO4, and GAL4, CPC1 of *Neurospora crassa*, and the mammalian transcription factors Sp1, CTF1, and C/EBP, and in a variety of nuclear protooncogene products (*myc, fos, jun, ets, myb*). These regulatory proteins may function by way of interactions with DNA, RNA polymerase, and/or other transcriptional factors. It appears that charge clusters are important for these functions (1, 2, 24). E1A and large-T antigen were recently shown to form complexes with RB (14, 19). RB also contains a significant acidic charge cluster, consistent with the predominance of special charge configurations in regulatory proteins discussed further below.

In terms of charge configurations, polymerase polypeptides are of diverse character. Thus the DNA polymerase of the herpesviruses HSV1 and VZV contains significant charge clusters that are not manifest in the homologues of EBV and CMV (7). The vaccinia DNA polymerase subunit is the only protein of 50 available of this virus containing a significant charge cluster (see Table 2); the RNA/DNA polymerase of influenza A contains a positive charge cluster; the reverse transcriptase of hepatitis B and HIV features periodic charge patterns, and the same holds for the adenovirus DNA polymerase. Interestingly, also the σ subunit of *E. coli* RNA polymerase and the 215K σ subunit of RNA/DNA-directed polymerase II of yeast contain charge clusters (data not shown). It is intriguing to speculate that these special charge structures may in some sense be complementary to the charge patterns of transcription factors.

Of some 150 proteins of prokaryotic phages, primarily tail components of λ and a tail protein of phage PZA of *B. subtilis* contain significant charge clusters (see Table 4). Is it possible that the charge clusters in the tail proteins of these phage facilitate attachment to suitable receptors on the host cell wall? In animal viruses capsid proteins or glycoproteins appear to interact with receptors on animal cell membranes, allowing the virions to be engulfed and then uncoated inside the cell. Do charge structures play a role in these processes?

We noted above the marked contrast in the proportions of significant charge structures among proteins of eukaryotic double-stranded DNA viruses compared with rather scarce charge structures among proteins of eukaryotic RNA viruses. Why this disparity? Eukaryotic DNA viruses (except poxviruses) replicate, transcribe, process mRNA, assemble and package virions in the nucleus with access to the cell's repertoire of polymerase activities, transcription factors, splicing and poly(A)-adding enzymes. Various DNA viruses can infect nondividing cells (e.g., cytomegalovirus, adenovirus, papovavirus) and stimulate cells to initiate DNA synthesis. Thus, the virus competes with, alters, represses, or activates transcription of the host DNA, expropriating cellular machinery for its own ends. DNA viruses entail early and late and in some cases (herpesviruses, adenoviruses) even multiple phases of gene expression. Electrostatic gradients are thought to play a role in protein sorting, transport, docking, and localization. Presumably, the charge structures in the regulatory proteins pertain to all of these processes. By contrast with DNA viruses, RNA viruses (excepting influenza viruses and retroviruses) replicate and mature in the cytoplasm, requiring few regulatory steps and modest association with cellular vesicles and processes.

Further data for >1500 protein sequences for a variety of eukaryotic and prokaryotic species revealed a marked correlation between nuclear regulatory proteins and multiple significant charge structures (not shown). In a collection of about 250 distinct *E. coli* proteins, only 9 cases of charge clusters were detected. On the other hand, from 300 human proteins of broad scope, about 25 sequences with significant charge clusters were discerned occurring principally among transcription factors and protooncogene products, (steroid) hormone receptors, and voltage-gated ion channel polypeptides. Thus, the set of *E. coli* gene products contains sharply fewer proteins with charge clusters (about 4%) compared to the set of mammalian polypeptides (about 8%). Examination of about 150 *Drosophila* and 110 yeast proteins revealed in each of these species ≈12–15% of sequences containing significant charge clusters, mostly heat shock proteins, developmental regulatory proteins, and transcription factors.

All of the foregoing indicates that significant charge configurations in proteins associated with regulatory function are the predominant theme. The charge structures identified in this paper should be attractive targets for experimental manipulation.

We thank Drs. R. Baldwin, E. Blaisdell, A. Campbell, and E. Mocarski for helpful discussions. This work was supported in part by National Institutes of Health Grants GM10452-26 and GM39907-01 to S.K. and by Sloan Foundation Grant B1987-2 to V.B.

1. Ma, J. & Ptashne, M. (1987) *Cell* **48**, 847–853.
2. Hope, I. A., Mahadevan, S. & Struhl, K. (1988) *Nature (London)* **333**, 635–640.
3. Kayano, T., Noda, M., Flockerzi, V., Takahashi, H. D. & Numa, S. (1988) *FEBS Lett.* **228**, 187–194.
4. Tempel, B. L., Jan, Y. N. & Jan, L. Y. (1988) *Nature (London)* **332**, 837–839.
5. Kalderon, D., Roberts, B. L., Richardson, W. D. & Smith, A. E. (1984) *Cell* **39**, 499–509.
6. von Heijne, G. (1986) *J. Mol. Biol.* **192**, 287–290.
7. Karlin, S., Blaisdell, B. E., Mocarski, E. S. & Brendel, V. (1988) *J. Mol. Biol.*, in press.
8. Blaisdell, B. E. & Karlin, S. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6637–6641.
9. Karlin, S., Ost, F. & Blaisdell, B. E. (1988) *Patterns in DNA and Amino Acid Sequences and their Statistical Significance*, ed. Waterman, M. (CRC, Boca Raton, FL), Chap. 6, in press.
10. Shenk, T. (1989) in *The Oncogenes*, eds. Wigler, M. & Weinberger, R. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), in press.
11. Ralston, R. & Bishop, J. M. (1983) *Nature (London)* **306**, 803–806.
12. Lillie, J. W., Loewenstein, P. M., Green, M. R. & Green, M. (1987) *Cell* **50**, 1091–1100.
13. Moran, E. & Mathews, M. B. (1987) *Cell* **48**, 177–178.
14. Whyte, P., Buchkovich, K. J., Horowitz, J. M., Friend, S. H., Raybuck, M., Weinberg, R. A. & Harlow, E. (1988) *Nature (London)* **334**, 124–129.
15. Horwitz, M. S. (1985) in *Virology*, eds. Fields, B. N., Knipe, D. M., Melnick, J. L., Chanock, R. M., Roizman, B. & Shope, R. E. (Raven, New York), pp. 433–476.
16. Sung, M. T., Cao, T. M., Coleman, R. T. & Budelier, K. A. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2902–2906.
17. Labow, M. A. & Berns, K. I. (1988) *J. Virol.* **62**, 1705–1712.
18. Christiansen, G., Landers, T., Griffith, J. & Berg, P. (1977) *J. Virol.* **21**, 1079–1084.
19. DeCaprio, J. A., Ludlow, J. W., Figge, J., Shew, J.-Y., Huang, C.-M., Lee, W.-H., Marsilio, E., Paucha, E. & Livingston, D. M. (1988) *Cell* **54**, 275–283.
20. Phelps, W. C., Yee, C. L., Münger, K. & Howley, P. M. (1988) *Cell* **53**, 539–547.
21. Goff, S. P. (1984) *Curr. Top. Microbiol. Immunol.* **112**, 45–71.
22. Gardner, R. C., Howarth, A. J., Hahn, P., Brown-Luedi, M., Shepherd, R. J. & Messing, J. (1981) *Nucleic Acids Res.* **9**, 2871–2888.
23. Garcea, R. L., Salunke, D. M. & Caspar, D. L. D. (1987) *Nature (London)* **329**, 86–87.
24. Gill, G. & Ptashne, M. (1987) *Cell* **51**, 121–126.