Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 4832–4836, July 1988
Immunology

Molecular cloning of a cDNA encoding the human Sm-D autoantigen

(autoimmunity/small nuclear ribonucleoprotein/λgt10 library/Epstein–Barr nuclear antigen/systemic lupus erythematosus)

LUIS A. ROKEACH, JEANNE A. HASELBY, AND SALLIE O. HOCH

The Agouron Institute, 505 Coast Boulevard South, La Jolla, CA 92037

Communicated by John N. Abelson, March 14, 1988

ABSTRACT **Antibodies to the Sm-D polypeptide antigen are closely associated with the rheumatic disease systemic lupus erythematosus. Sm-D exists in the cell as one of the core proteins of the small nuclear ribonucleoprotein complexes implicated in RNA processing. We have isolated a cDNA clone, D45-2, coding for the Sm-D human nuclear antigen by screening a human B-lymphocyte cDNA library with synthetic oligonucleotide probes. The 1633-base-pair clone contains an open reading frame (ORF) 357 nucleotides long, capable of encoding a 13,282-dalton polypeptide. The Sm-D coding region is initiated at an AUG codon downstream from a sequence with excellent match to the consensus for the eukaryotic ribosome-binding site. The Sm-D ORF is preceded by a 150-nucleotide-long untranslated leader and followed by a 1126-nucleotide-long untranslated region containing four putative poly(A) signals. The predicted amino acid sequence reveals a (Gly-Arg)$_9$ repeated motif at the C terminus, which may constitute one of the Sm-D immunoreactive determinants. Moreover, this C terminus shows interesting features: (*i*) a good homology to protamines as expected for a nucleic acid binding protein and (*ii*) a striking similarity to a region in the Epstein–Barr nuclear antigen.**

In autoimmune diseases, patients develop antibodies directed against normal cellular constituents (for review, see ref. 1). For instance, antinuclear antibodies with Sm specificity are a marker of systemic lupus erythematosus. The anti-Sm antibodies react with the group of small nuclear ribonucleoprotein complexes (snRNPs) associated with the U1, U2, U4, U5, and U6 small nuclear RNAs (snRNAs); anti-(U1) RNP sera, another immunospecificity often found along with anti-Sm, precipitate the U1 snRNP only (2). This finding means that the relevance of anti-Sm and anti-(U1) RNP antibodies is not restricted to autoimmune disease; indeed, these antibodies have provided valuable tools in the study of a basic cellular process—i.e., RNA splicing (for reviews, see refs. 3, 4). Sera with anti-Sm and anti-(U1) RNP specificities were used in *in vivo* and *in vitro* experiments to provide evidence of the Sm snRNP involvement in splicing of higher eukaryotic pre-mRNA.

In consequence, considerable effort has been drawn to the characterization of snRNPs as antigens and to the determination of their functional role during the splicing reaction. A core of at least six common polypeptides is shared by the Sm snRNPs: B′ (27 kDa), B (26 kDa), D (13 kDa), E/F (11-kDa doublet), and G (<10 kDa). In addition, there are unique polypeptides associated with individual snRNPs: 70 kDa, A (32 kDa), and C (18.5 kDa) belong to the U1 snRNP, whereas A′ (31 kDa) and B″ (26.5 kDa) are components of the U2 snRNP (5, 6). To date, the Sm antigenicity has been mainly associated with polypeptides B, B′, and D and to a lesser extent with E; (U1) RNP and (U2) RNP immunospecificities were assigned to their respective unique proteins (5, 7). It is thought that the unique proteins in each snRNP are essential elements for the specific interactions between snRNP/pre-mRNA and snRNP/snRNP. One current model of splicing suggests that the enzymatic activity of snRNP may be due solely to the RNA moiety, whereas the proteins would ensure an adequate configuration for catalysis and interaction between the different components of the splicing machinery (4). However, protein mediated catalysis cannot completely be ruled out as yet. The small core proteins were found to bind to a single-stranded sequence, A(U)$_n$G, where n = 3–6 (8), known as the Sm binding site, present in the U1, U2, U4, and U5 snRNAs. *In vivo* assembly experiments (9) indicate that proteins D to G first form an RNA-free complex to which B and B′ associate; this complex of proteins would bind subsequently to the U snRNAs. The D protein, which shows high affinity for RNA (10, 11), was proposed to mediate the binding of the core proteins to the snRNA. The determination of the amino acid sequence and the identification of immunoreactive epitopes in these snRNP proteins have capital importance in the elucidation of the cellular function and the role of these polypeptides in autoimmune disease pathogenesis. Here we report the isolation of a cDNA clone* containing the entire coding sequence of the Sm-D nuclear antigen and the analysis of the deduced amino acid sequence.

MATERIALS AND METHODS

Isolation of Sm-D Polypeptide. Sm snRNPs were isolated from HeLa extracts by immunoaffinity chromatography as described (12, 13). The individual snRNP polypeptides were fractionated by NaDodSO$_4$/polyacrylamide preparative gel electrophoresis. The polypeptides were isolated in microgram quantities by electroelution according to the protocol of Hunkapiller *et al.* (14), with modification of the staining procedure. Since a number of polypeptides were closely spaced, the Coomassie blue concentration was decreased to 0.05% from the original protocol to allow for clear visualization of bands before excision from the gel. After electroelution of individual bands, each was checked for homogeneity by silver stain and for immunoreactivity by protein blot (15, 16). The N-terminal sequence of electroeluted Sm-D was carried out at the Protein Chemistry Laboratory, Washington University, Saint Louis, employing a gas-phase peptide/protein microsequencer (Applied Biosystems).

Oligonucleotide Probes. Based on the N-terminal sequence, two 26-mer oligonucleotide probes were synthesized (see Fig. 1). To reduce the number of possible molecular species due to the degeneracy of the genetic code, the D1 probe was

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: snRNP, small nuclear ribonucleoprotein; snRNA, small nuclear RNA; EBNA, Epstein–Barr nuclear antigen.

*The sequence reported in this paper is being deposited in the EMBL/GenBank data base (IntelliGenetics, Mountain View, CA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03798).

designed using deoxyinosine at the third position of the most ambiguous codons (17). The D1 probe was consequently made as a mixture of 64 different oligonucleotides. For the D2 probe, all 4 nucleotides were included at the third position of redundant codons. Based on codon usage (18), we only considered cytosine at the first position of leucine triplets; thereby the total number of molecular species for the D2 probe was reduced to 1024. The probes were labeled to a specific activity of $\approx 1.2 \times 10^9$ cpm/μg employing [γ-^{32}P]-ATP (6000 Ci/mmol; 1 Ci = 37 GBq; New England Nuclear) and T4 polynucleotide kinase.

Screening. A λgt10 cDNA library made from human B-lymphocyte poly(A)$^+$ RNA was purchased from Clonetech Laboratories (Palo Alto, CA). The λgt10 library was plated ($\approx$5000 plaques per 85-mm plate) on *Escherichia coli* Y1090 (19); plaques were transferred in duplicate to nitrocellulose filters as described (20). Hybridization to the probes was performed in 0.9 M NaCl/90 mM sodium citrate, pH 8/5$\times$ Denhardt's solution (Denhardt's solution = 0.02% bovine serum albumin/0.02% Ficoll/0.02% polyvinylpyrrolidone); probes were added to 10^6 cpm/ml. The filters were washed in 0.9 M NaCl/90 mM sodium citrate for 2 hr with three changes of buffer. Hybridization with probe D1 was carried out at 45°C, and washes were done at 30°C; with probe D2, hybridization and washes were done at 45°C. Filters were exposed to autoradiography. Plaques identified as potential positives were purified by two successive rounds of hybridization.

DNA Sequencing. The *Eco*RI cDNA inserts from positive λ recombinants were cloned in the M13mp18 vector (21). The nucleotide sequence was determined by the dideoxy chain-terminator method (22) employing dATP[^{35}S] (New England Nuclear). To determine the nucleotide sequence of the D45-2 noncoding strand a series of overlapping deletions was made by the method of Dale *et al.* (23) using the Cyclone system (International Biotechnologies, New Haven, CT). The nucleotide sequence of the D45-2 coding strand was completed employing four oligonucleotide primers (25–30 nucleotides long) designed on the basis of the known sequence, and subcloning different overlapping restriction fragments into M13mp18 (sequencing strategy, Fig. 2).

Expression of the Cloned D45-2 cDNA *in Vitro*. The 1633-base *Pst* I–*Eco*RI 45-2 cDNA fragment spanning the Sm-D coding sequence was inserted into the pTZ19R vector (24). The resulting plasmid was linearized with *Eco*RI and transcribed by using T7 RNA polymerase. Four micrograms of *in vitro* synthesized mRNA was translated by using a rabbit reticulocyte system (Promega Biotec, Madison, WI) and [^{35}S]methionine (Amersham). Immunoprecipitation was carried out essentially as described (20) with modifications in the buffers. The immunoprecipitation buffer was 50 mM Tris·HCl, pH 7.5/150 mM NaCl/0.5% Nonidet P-40/2 mM EDTA/0.3% aprotinin. Protein A-Sepharose beads were washed four times with Tris·HCl, pH 7.5/150 mM NaCl/0.05% Nonidet P-40/2 mM EDTA/0.2 mM phenylmethylsulfonyl fluoride. Translation products and immunoprecipitates were analyzed by NaDodSO$_4$/PAGE and fluorography.

```
      Met Lys Leu Val Arg Phe Leu Met Lys

5'    ATG AAA TTN GTN CGN TTT TTN ATG AAA   3'  coding
        G C       A     C C         G         strand

3'    TAC TTT GAI CAI GCI AAG GAI TAC  TT   5'  D1 probe
          C   C   C   A   C

3'    TAC TTT GAA CAA GCA AAA GAA TAC  TT   5'  D2 probe
          C   T   T   T   G   T
              G   G   G       G
              C   C   C       C
```

FIG. 1. Oligonucleotide probes used to screen a B-lymphocyte cDNA library for the Sm-D gene. Probes were designed based on the N-terminal amino acid sequence of the protein.

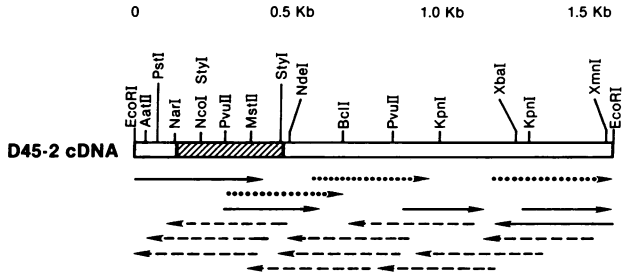

FIG. 2. Restriction map and sequence strategy of the D45-2 cDNA *Eco*RI fragment. Relevant restriction sites are indicated. The shaded area represents the Sm-D coding region. The DNA sequence was determined by the dideoxy chain-terminator method (22) using single-stranded DNA templates in the M13mp18 vector (21). The complete nucleotide sequence of D45-2 was obtained from restriction fragments (solid arrows), serial overlapping deletions (dashed arrows), and synthetic oligonucleotide primers (dotted arrows). Arrowheads show the direction of sequencing.

RESULTS

N-Terminal Sequence of the Sm-D Polypeptide. The approach that was eventually successful for us to purify the individual polypeptides of the Sm snRNP was to fractionate these peptides by NaDodSO$_4$/PAGE and electroelute. The protocol (and apparatus) that we used was that of Hunkapiller *et al.* (14) for the isolation of microgram amounts of protein. After elution, each polypeptide was checked for homogeneity by silver stain and for immunoreactivity by protein blot; the latter proved to be the more sensitive, detecting secondary bands not seen by silver stain. Electroeluted Sm-D was subjected to gas-phase amino acid sequencing, and the following 11-residue N-terminal sequence was obtained: Met-Lys-Leu-Val-Arg-Phe-Leu-Met-Lys-Leu-Ser.

Isolation of a cDNA Encoding the Sm-D Antigen. Approximately 250,000 independent λgt10 recombinants from the human B-lymphocyte library were plated and transferred to nitrocellulose filters. The filters were screened by hybridization with the ^{32}P-radiolabeled D1 probe (see Fig. 1). From this first round of screening, 60 potentially positive plaques were identified. The nitrocellulose filters were subsequently stripped of the D1 probe and rescreened with the ^{32}P-radiolabeled D2 oligonucleotide. By using this latter probe, 23 putative positive λ recombinants were obtained. Fifteen of these 23 clones hybridized previously with the D1 probe. All 23 positive recombinants were subjected to two more cycles of purification employing the D2 probe, thereby obtaining 18 pure positive recombinants. DNA from the positive λ clones was prepared (20) and digested with *Eco*RI, and the sizes of the cDNA inserts were determined by agarose gel electrophoresis. The inserts were classified into seven size groups ranging from 0.65 to 2.5 kilobases (kb). One representative from each size group was subcloned into M13mp18 and the nucleotide sequence was determined. The positive clone D45-2 contained a DNA stretch perfectly matching the codon composition for the known 11-amino acid sequence of the Sm-D protein.

DNA Sequence Analysis. The complete nucleotide sequence of D45-2 was determined and the 1633-base-pair sequence so obtained is shown in Fig. 3. Commencing at nucleotide 151 is an open reading frame (ORF) that begins with the nucleotides coding for the known amino acid sequence for the N terminus of the Sm-D autoantigen. This ORF has the capacity to encode a 119-residue polypeptide with a predicted molecular mass of 13,282 daltons. The first

```
                                        gaattccccc ccccccccccAGTGCTCCGCG    30
CGCTCTTGACGTCCGGAGCC CCTGGAGTAGGCGCTTCCGG CCATTCATACTGCAGTCGGT    90
CAGTGTTCGGTTGAAGGATT CTGTGTGCTGTCGGACCCAG AGGGTGACGGCGCCGCTAGG   150

   ATG AAG CTC GTG AGA TTT TTG ATG AAA TTG AGT CAT GAA ACT GTA ACC 198
 1 MET LYS LEU VAL ARG PHE LEU MET LYS LEU SER His Glu Thr Val Thr

   ATT GAA TTG AAG AAC GGA ACA CAG GTC CAT GGA ACA ATC ACA GGT GTG 246
17 Ile Glu Leu Lys Asn Gly Thr Gln Val His Gly Thr Ile Thr Gly Val

   GAT GTC AGC ATG AAT ACA CAT CTT AAA GCT GTG AAA ATG ACC CTG AAG 294
33 Asp Val Ser Met Asn Thr His Leu Lys Ala Val Lys Met Thr Leu Lys

   AAC AGA GAA CCT GTA CAG CTG GAA ACG CTG AGT ATT CGA GGA AAT AAC 342
49 Asn Arg Glu Pro Val Gln Leu Glu Thr Leu Ser Ile Arg Gly Asn Arg

   ATT CGG TAT TTT ATT CTA CCA GAC AGT TTA CCT CTG GAT ACA CTA CTT 390
65 Ile Arg Tyr Phe Ile Leu Pro Asp Ser Leu Pro Leu Asp Thr Ile Arg

   GTG GAT GTT GAA CCT AAG GTG AAA TCT AAG AAA AGG GAA GCT GTT GCA 438
81 Val Asp Val Glu Pro Lys Val Lys Ser Lys Lys Arg Glu Ala Val Ala

   GGA AGA GGC AGA GGA AGA GGA AGA GGA AGA GGA CGT GGC CGT GGC AGA 486
97 Gly Arg Gly Arg Gly Arg Gly Arg Gly Arg Gly Arg Gly Arg Gly Arg

   GGA AGA GGG GGT CCT AGG CGA TAA TGTCTCTCAA GATTTCAAAGTCATATGAGA 540
113 Gly Arg Gly Gly Pro Arg Arg  *

TTTGGGATATTTTTTGTACA GGTTGTGTTTGTTTATGTCA GTTTTTAATAAACATAAATG   600
TGGGACAGAGCTGTCTATTT AGTATATCAAAGTTTTAGTA GTTTCCTCCACATTCACGAA   660
ATTACCACAGTGAGAGCTAA GCATTTCTACTGGGCAGTTT CATTTTTAGTTGATCAGGTT   720
TTAAGTTTTTGAACTAAAAT TTTTCTTTTTCTTTTTATGA TGAATAAGGTTAAAATAAAA   780
GCCTTAGACAAATTAAATTT GGCAGAGTTTAATTGAGCAA AGGACAATTCACAAATCAGG   840
TAGCCCCTGAACCATAATAG GCTCAGAGGCTTCAGCCCAG CTGCATAGTTGAAGATTTAT   900
GGACAGAAGGAAAGTGATGT ATGGAAAATGGAAGTGAGAT ACAGCAACAGCCGGATTAGT   960
TACAGTTCAGCGTTTGCCTT ATTTGAATATGGTTTGAACA GTTCGCTGTCTTTGGTTGGC  1020
TGAAACTTAGTGATTGCCAC AAGAGTAGGGTACCGTCTGT TTACACGTCCAGTTAGGCTA  1080
CAGTTCTATGTACTGAGAAA CCTTTAAGCTGAACTTGAGA TATGTAAAGAGACTTTAGGC  1140
TAAACTTAACAATATATATA GGAATATATCCCTTCTACTT CACATGCACTGAATATGCAT  1200
TTTATTGCTTTACTCTTCAT TCTGTGGCACCTACCCACAG GGGAAGTAAGAAGTTTGTTT  1260
TGGTATTTCGGAAACTAAAG TCCTTATGGGATGGGGTCTA GAATTGATTCTCCTTTCCTG  1320
AGTTTTACTCCACGGAGTCT TAGGTACCTGGTAAAAAGTT GTCTTCTAAATTAAGGGTCA  1380
TTGCTTTGTTGTCTAGCTGC TAATGTCTTACTTTTGTTTC TTTTGCTTTTTAATCAGTTC  1440
TTAATAGGATATAGTTTTAT GTTTTCCAAGTTATAACTTG GAGTTAATGGTCACTAGATT  1500
ATCAGTTATGAGCAGTGTTA AAATCTCCTATTAATGTGTA ATGTACCTGTCAGTGCCTCC  1560
TTTATTAAGGGGTTCTTTGA GAATAAAAGAGAAAAGACCT ACTTTATTTGACAGCAAAAA  1620
AAAAAAggaattc                                                    1633
```

FIG. 3. Nucleotide and deduced amino acid sequences of D45-2 cDNA. Extraneous sequences added during cDNA synthesis and cloning processes are shown in lowercase characters. Amino acids indicated in uppercase letters were determined by sequencing of the N-terminal part of the protein and used in the design of oligonucleotide probes. Nucleotides are numbered on the right; amino acids are numbered on the left. The possible 18S rRNA-binding site (25) is underlined. Doubly underlined nucleotides represent potential poly(A) signals (26). The termination codon TAA is indicated by an asterisk.

AUG of this reading frame is located in an environment propitious for translation: (*i*) it is the first AUG triplet on the cDNA clone; (*ii*) there is an adenine residue 3 nucleotides upstream from the AUG codon; this is also a well-conserved feature of most eukaryotic messages (27); and (*iii*) the sequence 5′ CCGCTAG 3′ located between nucleotides 143 and 149 has an excellent complementarity to the sequence 5′ CCUGCGG 3′ found at the 3′ end of the 18S rRNA (25). It is therefore very likely that translation of the Sm-D protein is initiated at the AUG codon located between nucleotides 151 and 153 and that no *in vivo* processing occurs at the N terminus of the polypeptide. Following the TAA termination triplet there is a 1126-nucleotide-long untranslated region containing several putative poly(A) signals (26) at positions 587–592 (AATAAA), 774–779 (AATAAA), and 791–797 (AATTAAA) and the one that probably was used in this clone, 1582–1587 (AATAAA). This multiplicity of potential poly(A) signals may result in a heterogeneity of mRNA sizes.

***In Vitro* Translation and Immunoprecipitation.** D45-2 mRNA obtained by *in vitro* transcription was used in a rabbit reticulocyte translation system employing [^{35}S]methionine. A sample of the translation reaction was subjected to direct analysis. Another portion of the translation assay was immunoprecipitated by using affinity-purified human anti-Sm IgG or normal human antibodies as control. The radiolabeled proteins were visualized by fluorography after NaDodSO$_4$/PAGE (Fig. 4). Two prominent labeled species were observed, one comigrating with unlabeled Sm-D from human snRNP, the other ≈10% larger. Both polypeptides were immunoprecipitable by human anti-Sm antibodies, indicating that these proteins are products of the same gene. The same results were observed after immunoprecipitation with two other human anti-Sm antibodies and a murine monoclonal antibody (7.13) that recognizes the Sm-D polypeptide (16).

Predicted Secondary Structure of Sm-D. The Chou–Fasman algorithm (28) predicts (Fig. 5) two regions of highly hydrophilic character at the C-terminal half of the protein. One is a cluster rich in lysine located between residues 86 and 92. The other hydrophilic domain is a stretch of nine alternating Gly-Arg repeats that have the potential to form a series of β-turns induced by the glycine residues. The composition of these two juxtaposed domains would suggest that the hinge region between the domains might be readily susceptible to

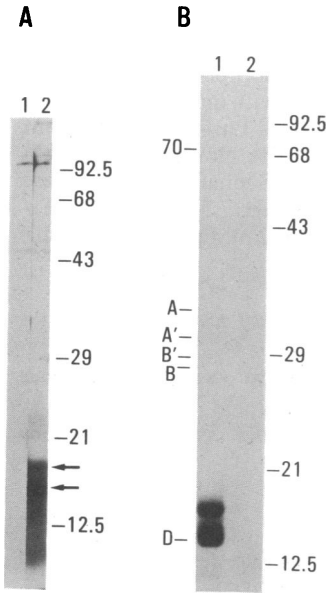

FIG. 4. *In vitro* translation and immunoprecipitation of Sm-D encoded by the D45-2 cDNA. The D45-2 cDNA was subcloned into an expression vector and transcribed *in vitro* with T7 RNA polymerase. (*A*) Lane 1, translation with no exogenous RNA; lane 2, synthetic D45-2 mRNA translated *in vitro* employing a rabbit reticulocyte lysate system and [^{35}S]methionine. (*B*) ^{35}S-radiolabeled proteins immunoprecipitated with human anti-Sm IgG (lane 1) or with normal human antibodies (lane 2). Radiolabeled polypeptides were fractionated by NaDodSO$_4$/PAGE and visualized by fluorography. Positions of Coomassie blue-stained size markers (in kDa) are shown on the right in *A* and *B*, and positions of stained HeLa Sm snRNP proteins are given on the left in *B*. Arrows indicate positions of the major *in vitro* translation products.

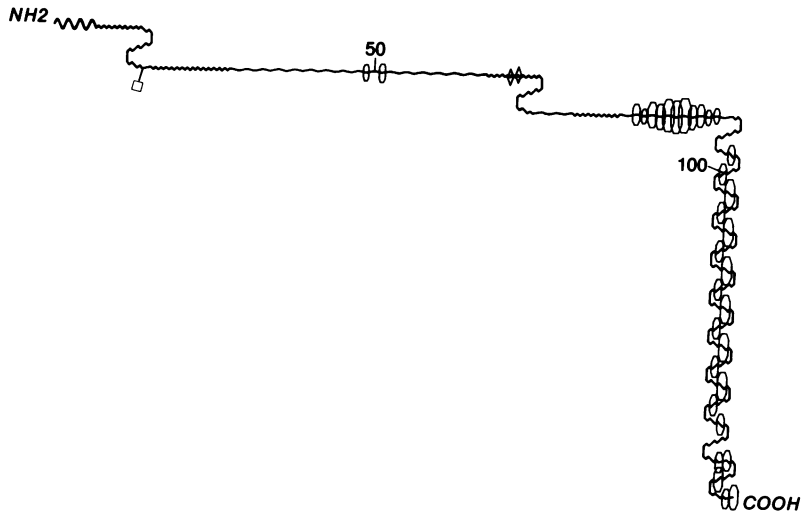

FIG. 5. Chou–Fasman plot (28) of the predicted structure of Sm-D. The hydrophobic moment (diamond) and charged residues (octagon) are indicated.

proteolytic attack. In fact, amino acid analysis of electroeluted Sm-D polypeptide infers the loss of the terminal tail region.

DISCUSSION

We have identified a cDNA clone, D45-2, that encodes the human Sm-D autoantigen. Our cloning approach consisted of screening a human B-lymphocyte cDNA library employing radiolabeled oligonucleotide probes designed after the amino acid sequence of the N-terminal portion of the Sm-D polypeptide. Although at first glance the immunoscreening of an expression library seems more reasonable, and various groups have succeeded in the cloning of cDNAs encoding nuclear antigens (29–33) by using this approach, we opted to screen with oligonucleotide probes for the following reasons: (*i*) human serum is known for having multiple antibody specificities and (*ii*) even monoclonal antibodies may recognize the same epitope shared by several proteins, making it difficult to prove unequivocably the identity of the protein. Therefore, we took advantage of our ability to isolate snRNP proteins in amounts large enough for peptide sequencing. The 11 amino acids at the N terminus deduced from DNA analysis are identical to those obtained by protein sequencing. This indicates that translation starts at the designated AUG codon (nucleotide positions 151–153) and no *in vivo* posttranslational processing occurs at the N-terminal part of the polypeptide. The most salient feature of the deduced amino acid sequence of Sm-D is a 9-fold Gly-Arg repeat located at the C terminus of the polypeptide. A Chou–Fasman analysis (28) indicates that the Gly-Arg repeats have the potential to generate a series of turns preceded by a strongly hydrophilic region rich in lysine. It is therefore likely that these two regions of the protein are exposed to the surface and constitute one or more of the Sm-D antigen epitopes. Antibodies to a synthetic $(Gly\text{-}Arg)_9$ polypeptide may be generated to assist in the characterization of this repeat.

```
Sm-D (89-119)          SKKREAVAGRGRGRGRGRGRGRGRGGPRR
                       •••| | •|| | | • ••| | |||•||
Protamine I (1-31)     PRRRRASSGRPVRRRRRPKMSRRRRRGGRRR

Sm-D (96-119)          AGRGRGRGRGRGRGRGRGGPRR
                       •| ••|||||||||||| || |•
EBNA (35-58)           GGDNHGRGRGRGRGRGGGRPGAPG
```

FIG. 6. Comparison of the predicted amino acid sequence (in the single-letter code) of Sm-D with protamine I from northern pike (35) and Epstein–Barr nuclear antigen, EBNA-1 (36). Identities are indicated by bars; conservative changes are denoted by dots; in parentheses are the range of residues aligned.

We have addressed the question of immunoreactivity of the recombinant Sm-D polypeptide by *in vitro* translation. *In vitro* synthesized D45-2 mRNA directed the synthesis of two polypeptides, one that comigrated with Sm-D and one slightly larger, both of which were immunoprecipitable with anti-Sm antibodies. The smaller polypeptide might represent a major proteolytic product of the larger one, as implied by the background smear seen in the direct analysis of the *de novo* synthesized proteins. This observation leads to the suggestion that the larger product might be the intact Sm-D polypeptide that would migrate anomalously relative to its predicted molecular mass. The smaller protein could arise from the loss of a large oligopeptide, possibly the $(Gly\text{-}Arg)_9$ tail. Consequently, this suggestion implies that the $(Gly\text{-}Arg)_9$ repeat may be lost during purification of the snRNP proteins, as implied by amino acid analysis of the electroeluted polypeptide. An alternative explanation for the presence of two radiolabeled products is that the larger rises as the result of inefficient termination at the TAA stop codon under the *in vitro* conditions.

Because the deduced Sm-D amino acid sequence has segments particularly rich in basic residues, this protein is expected to have a high binding affinity for nucleic acids, as was experimentally observed (10, 11). A computer search for similarities (34) to other proteins (in the National Biomedical Research Foundation data base†) showed alignments with several piscine protamines; an example is shown in Fig. 6. It is not surprising that a protein rich in lysine and arginine would have structural similarities to protamines; therefore, the significance of this similarity may be limited. However, the C terminus also has a strong similarity to a well-conserved protamine sequence—namely, Arg-Gly-Gly-Arg-Arg-Arg (37). Protamine-like domains were also found in the 70-kDa protein from U1 snRNP (30, 33). Recently, a conserved amino acid stretch, denoted as the RNP consensus, was found in poly(A)-binding proteins, heterogeneous nuclear ribonucleoproteins (hnRNP), and several single-stranded RNA- and DNA-binding proteins (38). It was therefore suggested that the RNP consensus sequence forms part of a larger RNA-binding domain. No similarities to the RNP consensus were found in the deduced amino acid sequence of

†Protein Identification Resource (1987) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 13.0.

Sm-D. Nevertheless, as discussed by Swanson *et al.* (38), the RNP consensus is not necessarily a general property of RNA-binding proteins since it is absent in ribosomal proteins and in many capsid proteins in RNA viruses. It is intriguing that the RNP consensus is found in the snRNP unique polypeptides, 70 kDa and B″ (30, 31, 33), but is absent in the snRNP core proteins, Sm-D, Sm-E (39), and Sm-B/B′ (unpublished data). It can be speculated that the core snRNP proteins, lacking the RNP consensus sequence, may have general scaffold functions, facilitating nonspecific RNA–protein interactions through their basic domains and/or facilitating snRNP–snRNP or snRNP–hnRNP interactions within the spliceosome.

The best alignment we have found may have more relevance to Sm-D as an autoimmune antigen than as a snRNP component; this alignment again encompasses the Gly-Arg repeat. This similarity is with EBNA-1, which is produced in either lytic or lysogenic EB infection (Fig. 6). It is yet unclear how systemic lupus erythematosus and other autoimmune diseases arise. One of the current theories, known as the *molecular mimicry* hypothesis, is that as the first step infecting agents trigger the generation of antibodies; this primary immune response is perpetuated subsequently by autoantigens sharing epitopes with the foreign antigens (40). The Gly-Arg repeat does not correspond to either of the two major epitopes that have been identified for EBNA-1, one in the internal Gly-Ala repeat and the second at the C terminus of the protein (41–44). However, it is possible that the Gly-Arg region in EBNA-1 normally remains occluded inside the molecule, but in a small number of EBV carriers the Gly-Arg region becomes exposed as a result of proteolytic degradation or change in configuration of the protein. Thereby, the first step in the pathogenesis of systemic lupus erythematosus would take place.

We are grateful to Kathleen Trach and James Hoch for their valuable tutelage, advice, and encouragement and to Richard Ogden for his support. We thank Leticia Wolpert and La Vonne Melheim for their assistance in typing this manuscript. This work was supported by Research Grants AI21083, AR36387 (to S.O.H.), and GM19416 (to J.A.H.) from the National Institutes of Health.

1. Sharp, G. C. & Alspaugh, M. (1985) in *Immunology of Rheumatic Diseases*, eds. Gupta, S. & Talal, N. (Plenum, New York), pp. 197–219.
2. Lerner, M. R. & Steitz, J. A. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5495–5499.
3. Sharp, P. A. (1987) *Science* **235**, 766–771.
4. Maniatis, T. & Reed, R. (1987) *Nature (London)* **325**, 673–678.
5. Brunel, C., Sri-Widada, J. & Jeanteur, P. (1985) *Prog. Mol. Subcell. Biol.* **9**, 1–52.
6. Bringmann, H. & Lührmann, R. (1986) *EMBO J.* **5**, 3509–3516.
7. Habets, W., Hoet, M., Bringmann, P., Lührmann, R. & van Venrooij, W. (1985) *EMBO J.* **4**, 1545–1550.
8. Liautard, J. P., Sri-Widada, J., Brunel, C. & Jeanteur, P. (1982) *J. Mol. Biol.* **162**, 623–643.
9. Fisher, D. E., Conner, G. E., Reeves, W. H., Wisniewolski, R. & Blobel, G. (1985) *Cell* **42**, 751–758.
10. Tazi, J., Alibert, C., Temsamani, J., Reveillaud, I., Cathala, G., Brunel, C. & Jeanteur, P. (1986) *Cell* **47**, 755–766.
11. Gerke, V. & Steitz, J. A. (1986) *Cell* **47**, 973–984.
12. Billings, P. B. & Hoch, S. O. (1984) *J. Biol. Chem.* **259**, 12850–12856.
13. White, P. J., Billings, P. B. & Hoch, S. O. (1982) *J. Immunol.* **128**, 2751–2756.
14. Hunkapiller, M. W., Lujan, E., Ostrander, F. & Hood, L. E. (1983) *Methods Enzymol.* **91**, 227–236.
15. Towbin, H., Staehelin, T. & Gordon, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4350–4354.
16. Billings, P. B., Barton, J. R. & Hoch, S. O. (1985) *J. Immunol.* **135**, 428–432.
17. Ohtsuka, E., Shigeru, M., Ikehara, M., Takahashi, Y. & Matsubara, K. (1985) *J. Biol. Chem.* **260**, 2605–2608.
18. Lathe, R. (1985) *J. Mol. Biol.* **183**, 1–12.
19. Young, R. A. & Davis, R. W. (1983) *Science* **222**, 778–782.
20. Davis, L. G., Dibner, M. D. & Battey, J. F. (1986) *Basic Methods in Molecular Biology* (Elsevier, New York).
21. Yanisch-Perron, C., Vieira, J. & Messing, J. (1985) *Gene* **33**, 103–119.
22. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
23. Dale, R. M. K., McClure, B. A. & Houchins, J. P. (1985) *Plasmid* **13**, 31–40.
24. Mead, D. A., Szczesna-Skorupa, E. & Kemper, B. (1986) *Protein Eng.* **1**, 67–74.
25. Chan, Y. L., Guttell, R., Noller, H. F. & Well, J. G. (1984) *J. Biol. Chem.* **259**, 224–230.
26. Birnstiel, M. L., Busslinger, M. & Strub, K. (1985) *Cell* **41**, 349–359.
27. Kozak, M. (1986) *Cell* **44**, 283–292.
28. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47**, 45–148.
29. Chambers, J. C. & Keene, J. D. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2115–2119.
30. Thiessen, H., Etzerodt, M., Reuter, R., Scheider, C., Lottspeich, F., Argas, P., Lührmann, R. & Phillipson, L. (1986) *EMBO J.* **5**, 3209–3217.
31. Habets, W. J., Sillekens, P. T. G., Hoet, M. H., Schalken, J. A., Roebroek, A. J. M., Leunissen, J. A. M., van de Ver, W. J. & van Venrooij, W. J. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 2421–2425.
32. Earnshaw, W. C., Sullivan, K. F., Machlin, P. S., Cooke, C. A., Kaiser, D. A., Pollard, T. D., Rothfield, N. F. & Cleveland, D. W. (1987) *J. Cell Sci.* **104**, 817–829.
33. Spritz, R. A., Strunk, K., Surowy, C. S., Hoch, S. O., Barton, D. E. & Francke, U. (1987) *Nucleic Acids Res.* **15**, 10373–10391.
34. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
35. Speckert, W., Kennedy, B., Daisley, S. L. & Davies, P. (1983) *Eur. J. Biochem.* **136**, 283–289.
36. Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. V., Hatfull, G., Hudson, G. S., Satchwell, S. C., Sequin, C., Tuffnell, P. S. & Barrell, B. G. (1984) *Nature (London)* **310**, 207–211.
37. Aiken, J. M., McKenzie, D., Zhao, H. Z., States, J. C. & Dixon, G. H. (1983) *Nucleic Acids Res.* **11**, 4907–4922.
38. Swanson, M. S., Nakagawa, T. Y., Le Van, K. & Dreyfuss, G. (1987) *Mol. Cell. Biol.* **7**, 1731–1739.
39. Stanford, D. R., Rohleder, A., Neiswanger, K. & Wieben, E. D. (1987) *J. Biol. Chem.* **262**, 9931–9934.
40. Oldstone, M. B. A. (1987) *Cell* **50**, 819–820.
41. Hennessy, K. & Kieff, E. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 5665–5669.
42. Rumpold, H., Rhodes, G. H., Bloch, P. L., Carson, D. A. & Vaughan, J. H. (1987) *J. Immunol.* **138**, 593–599.
43. Dillner, J., Sternås, L., Kallin, B., Alexander, H., Ehlin-Henrikson, B., Jörnvall, H., Klein, G. & Lerner, R. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4652–4656.
44. Milman, G., Scott, A. L., Cho, M., Hartman, S. C., Ades, D. K., Hayward, G. S., Ki, P., August, J. T. & Hayward, S. D. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 6300–6304.