Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 3963–3967, June 1986
Immunology

Murine protein H is comprised of 20 repeating units, 61 amino acids in length*

[complement protein H/full-length cDNA (sequence)/internal repeats/gene duplications]

TORSTEN KRISTENSEN AND BRIAN F. TACK

Division of Molecular Immunology, Research Institute of Scripps Clinic, La Jolla, CA 92037

Communicated by Frank W. Putnam, January 2, 1986

ABSTRACT **A cDNA library constructed from size-selected (>28 S) poly(A)$^+$ RNA isolated from the livers of C57B10.WR mice was screened by using a 249-base-pair (bp) cDNA fragment encoding 83 amino acid residues of human protein H as a probe. Of 120,000 transformants screened, 30 hybridized with this cDNA probe. Ten positives were colony-purified, and the largest plasmid cDNA insert, MH8 (4.4 kb), was sequenced by the dideoxy chain termination method. MH8 contained the complete coding sequence for the precursor of murine complement protein factor H (3702 bp), 100 bp of 5′-untranslated sequence, 448 bp of 3′-untranslated sequence, and a polyadenylylated tail of undetermined length. Murine pre-protein H was deduced to consist of an 18-amino acid signal peptide and 1216 residues of H-protein sequence. Murine H was composed of 20 repetitive units, each about 61 amino acid residues in length. Similar repetitive units are present in the C4b binding protein, the C3b-receptor (CR1), complement factor B and C2, and in β_2-glycoprotein I and the interleukin 2 receptor. This finding suggests a common evolutionary origin for regions of these proteins.**

Murine complement protein H, a plasma glycoprotein of β-mobility, is a cofactor for cleavage of fluid-phase C3b by the serine protease I (1–3); C3b is the major activation product of the third component of complement, C3. Protein H has also been studied in rabbits (4, 5) and humans (6–17) and is most fully characterized as a regulatory component of the alternative complement pathway in the latter species. Irrespective of the species of origin, protein H has been reported to have a M_r of 150,000–160,000 and a carbohydrate content between 4% and 18% by weight and to function as a cofactor for the conversion of C3b to iC3b by serine protease I. Protein H also has been shown to accelerate the dissociation of the Bb fragment of complement factor B from the alternative pathway C3 (C3bBb) and C5 ($C3b_2Bb$) convertases.

Two other proteins acting as cofactors for serine protease I-mediated cleavage of C4b (the major activation product of the fourth component of complement C4) and/or C3b have been described in humans and mice. One of these is the plasma protein C4b-binding protein (C4BP) that serves as a cofactor for conversion of C4b to iC4b (18–22), and the other is the C3b receptor (CR1 for complement receptor type 1). CR1 is an integral membrane glycoprotein of M_r 160,000–250,000 that acts as a cofactor for inactivation of both C4b and C3b by serine protease I (23–27).

Genetic studies have shown three codominant alleles for protein H in humans (28) and two in mice (3). In humans the loci for protein H, C4BP, and CR1 are closely linked (29). This may also be the case for the corresponding proteins in the mouse. We have shown that the structural gene for protein H is located on chromosome 1 in the mouse (P. D'Eustachio, T.K., R. A. Wetsel, R. Riblett, B. Taylor, and B.F.T., unpublished work). In humans, the CR1 structural gene has been mapped to the q region of chromosome 1 (31) for which there is an analogous region on chromosome 1 in the mouse (32).

We report here the complete coding sequence and deduced protein sequence for murine pre-protein H. These studies have indicated that protein H is composed of 20 repeating units, each about 61 amino acids in length. A repetitive unit of this nature has been previously observed in the Ba fragment of factor B (33, 34) and the C2b fragment of complement component C2 (35)—amino-terminal fragments produced upon cleavage of B and C2 by protein D and $\overline{C1}$, respectively—and in C4BP (22), CR1 (31), β_2-glycoprotein I (β_2-gpI) (36), and the interleukin 2 receptor (37, 38).

MATERIALS AND METHODS

A 249-base-pair (bp) *Sau*3A1 cDNA fragment encoding 83 amino acid residues of human protein H (39) was used as a probe in all hybridization experiments. For each hybridization, 1 μg of the cDNA probe was ^{32}P-labeled by nick-translation (40) using a commercially available nick-translation kit obtained from Bethesda Research Laboratories and α-^{32}P-labeled deoxycytidine and deoxyguanosine triphosphates from Amersham.

A cDNA library constructed by the procedures of Okayama and Berg (41, 42) from size-selected (>28 S) and methylmercury(II) hydroxide-denatured poly(A)$^+$ RNA isolated from the livers of C57B10.WR mice (43) was kindly made available to us by R. T. Ogata (Research Institute of Scripps Clinic). High-density screening of the cDNA library was performed on blotted duplicate nitrocellulose filters (44) as described (39, 45). Hybridizing colonies were removed with a toothpick and rescreened at a density of 200–300 colonies per filter. Single colonies were isolated and analyzed with the restriction enzymes *Hpa* I, *Xho* I, and *Bam*HI obtained from Boehringer Mannheim. Plasmids estimated to contain full-length or nearly full-length inserts were isolated by standard techniques (46). All inserts were excised in one piece by *Hpa* I (0.3 units/μg of DNA) and isolated from 0.9% agarose gels followed by extractions with phenol/chloroform and precipitation with ethanol.

cDNA sequences were determined by using the "shotgun" DNA sequencing strategy in M13 mp8 (47). Sonicated subfragments were produced from self-ligated cDNA inserts, inserted into the *Sma* I site of M13 and sequenced at random by the dideoxynucleotide technique (48) using α-[^{35}S]thio-

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: bp, base pair(s); kb, kilobase(s); C4BP, C4b binding protein; β_2-gpI, β_2-glycoprotein I; CR1, complement receptor type I (C3b receptor).

*Part of this work was presented at the Eleventh International Complement Workshop, November 3–5, 1985, in Key Biscayne, FL.

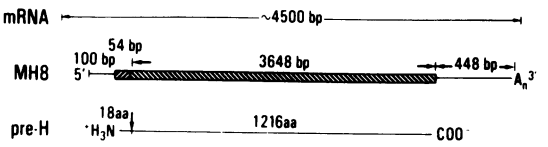

FIG. 1. (*Legend appears at the bottom of the opposite page.*)

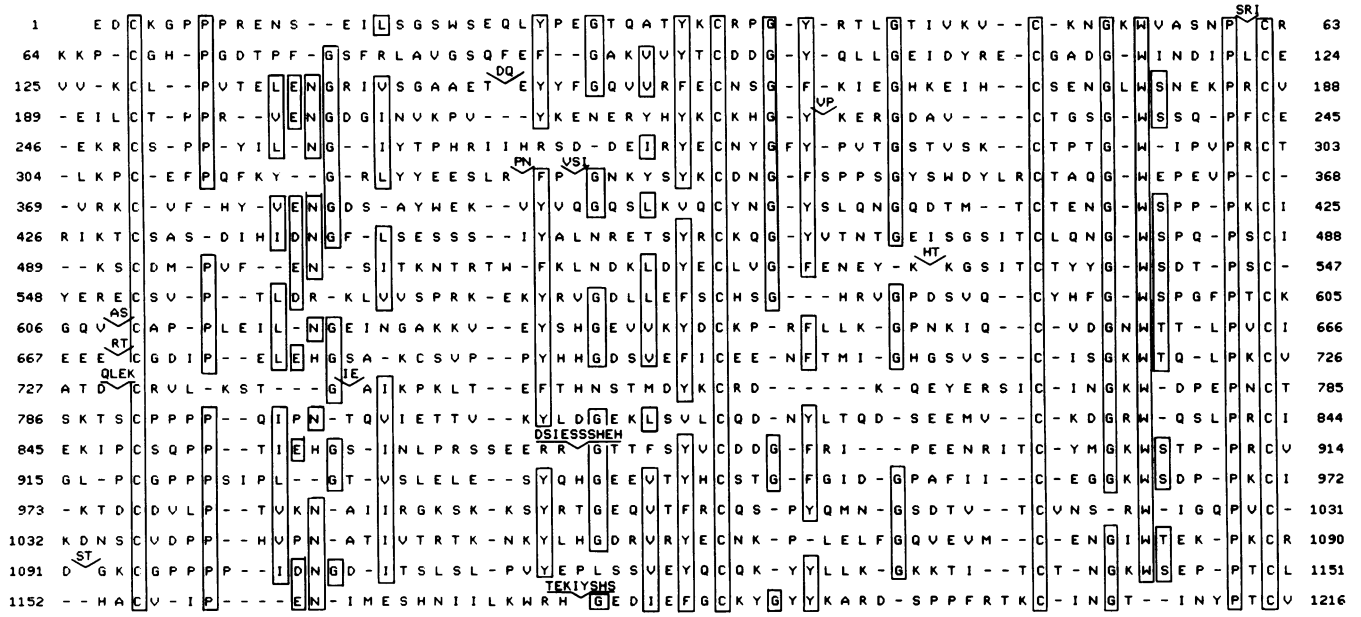

FIG. 2. Alignment of 20 regions of internal homology in the derived amino acid sequence of murine protein H. The single-letter amino acid code is used. Identical and similar residues (E=D, S=T, Q=N, F=Y, K=R=H) are boxed. The division into repetitive regions was done as described for the Ba fragment of factor B (57), where each region is encoded by a separate exon. Each line represents a repetitive unit.

labeled deoxyadenosine 5'-[thio]triphosphate (Amersham) and gradient gels (47, 49).

Sequence results were compiled and aligned by the DBAUTO and DBUTIL programs (50, 51). Sequence comparison analyses were performed with the graphics program DIAGON (52), the BESTFIT program of the University of Wisconsin Genetics Computer Group (Madison, WI), and by screening the National Biomedical Research Foundation (Georgetown University) protein sequence library.

RESULTS AND DISCUSSION

Isolation and Sequence Analysis of Murine Protein H cDNA Clones. A mouse liver cDNA library (120,000 transformants) was screened with a 249-bp ^{32}P-labeled cDNA probe encoding human protein H (39). Thirty hybridizing colonies were found. Ten of these were rescreened at lower density and purified. Initially, one of the largest candidate protein H cDNA clones [MH4, estimated to be about 4.0 kilobases (kb)] was fully sequenced. MH4 was confirmed as coding for murine protein H based on a derived protein sequence that exhibited strong homology (63% identically placed residues) with that of a previously characterized human protein H cDNA clone (39).

The amino acid sequence deduced from the 5' coding sequence of MH4 did not exhibit, however, the expected homology with the previously determined amino-terminal sequence of human protein H (39). On reexamination of the above isolated clones, MH8 (4.5 kb) was estimated to be about 500 bp longer at its 5' end and therefore likely to contain the coding sequence for the entire estimated pre-protein H molecule of 4.4 kb. The complete sequence for murine pre-protein H (Fig. 1) was deduced from sequence analysis of MH4 and MH8. The 5' end of MH8 was excised with *Xho* I at position 554 (Fig. 1). This fragment, which overlapped the 5' sequence of MH4, was sequenced to the poly(G) tail, thereby extending the protein H cDNA sequence by 201 bp. A 48-bp intron sequence (data not shown) was present in MH4 at position 3424 (Fig. 1). It contained a stop codon within an otherwise open reading frame. This intron was proven to be absent in MH8 by sequencing a 239-bp *Rsa* I fragment spanning positions 3333–3571 (see Fig. 1). Sequence analysis of random *Rsa* I fragments covering most of the MH8 structure confirmed the identity between MH8 and MH4. Thus, a contiguous sequence of 4252 bp was generated from sequence analysis of MH4 and MH8 cDNAs coding for murine protein H (Fig. 1). This sequence included 100 bp of the 5' untranslated region followed by 54 bp of coding sequence for an 18-amino-acid-long potential signal peptide sequence of hydrophobic character (54, 55). Downstream of the putative signal peptide sequence were 3648 bp of coding sequence for 1216 amino acids of protein H followed by a stopcodon (TAA), 448 bp of 3' untranslated region, and a polyadenylylated tail. The putative polyadenylylation recognition signal was A-A-T-T-A-A located 18 bp upstream from the polyadenylylated tail.

Based on the general sequence Asn-Xaa-Ser/Thr, there are eight potential sites for asparagine-linked carbohydrate chains (indicated by asterisks in Fig. 1). Due to the presence of a proline residue at position 58, the asparagine at position 57 is unlikely to be N-glycosylated (53).

Murine protein H has a high content of cysteine (81 residues; 6.7 mol %) and of proline (97 residues; 8.0 mol %), which has also been observed for rabbit and human proteins H (4, 13). CD spectral studies of human protein H have

FIG. 1 (*on opposite page*). Nucleotide coding sequence of murine protein H cDNA and derived amino acid sequence. The single-letter amino acid code is used. Potential glycosylation sites at asparagine residues are indicated by asterisks. The asterisk above position 57 is in parentheses, as this asparagine residue is not likely to be glycosylated because of the presence of proline in position 58 (ref. 53). The amino-terminal sequence (positions 1–33) and tryptic peptide sequences (positions 490–500, 626–643, 648–706, 837–864, 894–906, 907–945, 946–959, 1150–1158, and 1159–1169) of human protein H and deduced amino acid sequence of a partial human protein H cDNA (positions 208–431) are indicated by arrowheads (◄ = start and ► = stop, except for the amino terminus). Identically placed residues between human and murine protein H amino acid sequences are underlined.

indicated a very unusual conformation dependent on intact disulfide bonds and the absence of α-helical and β-sheet structures (17). It has been proposed that human protein H is an elongated molecule because of its high frictional coefficient and elution as a M_r 300,000 protein on gel filtration (7, 13). EM studies have indicated that human protein H has both a globular and an elongated rod-shaped domain (56). Such studies have not been performed on murine protein H; however, the high degree of sequence homology (63%) between known murine and human protein H sequences suggests that their higher-ordered structures will be similar.

Presence of a Repeating Unit Within the Deduced Protein H Sequence. Examination of the derived amino acid sequence of protein H revealed the presence of 20 regions that exhibited internal homology (Fig. 2). Each region was about 61 amino acids long. While significant sequence variability was apparent between the repeating units of protein H, the following amino acid residues were conserved (boxed in Fig. 2): 4 cysteines, 5 glycines, 3 tyrosines/phenylalanines, 2 prolines, 3 isoleucines/leucines/valines, 1 tryptophan, 1 asparagine, 1 serine/threonine, and 1 glutamic acid/aspartic acid. The amino-terminal ends of several of the repetitive units were also observed to be rich in proline.

Homology of the Protein H Repeating Unit with Other Protein Sequences. Similar regions of internal homology are present in the human complement proteins C4BP (22), CR1 (31), B (33, 34), and C2 (35), in human β_2-gpI (36), and in human and mouse interleukin 2 receptors (37, 38). C4BP, a regulatory protein of the classical-pathway C3 convertase (18–21), contains eight of these 61-amino-acid-long repeats and a carboxyl-terminal nonhomologous region (22). The homology between the repetitive regions of murine protein H and those of human C4BP is illustrated by DIAGON analysis in Fig. 3 *Top*. CR1, which functions as a control protein for both the classical- and the alternative-pathway C3 convertases (C4b2a and C3bBb, respectively) in a manner similar to protein H and C4BP, has recently been partially sequenced (31). This protein may be composed of 30 repetitive regions, 10 of which have been sequenced and shown to be homologous with those of murine protein H (data not shown). The loci for protein H, C4BP, and CR1 are linked in humans (29). The locus for the structural CR1 gene has been mapped to a single site on the long arm of chromosome 1 in humans (31). Recently, we have assigned the murine protein H locus (*Cfh*) to chromosome 1 (30). The *Cfh* locus is either identical with or adjacent to the previously described *Sas-1* locus. Because of the location of the murine CR1 structural gene on chromosome 1 (John Weis, personal communication), the close linkage of genes for protein H, C4BP, and CR1 in humans (29), and the fact that the q region of chromosome 1 in humans is analogous to the same region in mice (32), it is likely that the murine C4BP gene will also be located on this chromosome. Thus, three structurally, as well as functionally, similar glycoproteins may have originated from a common ancestral gene coding for a 61-amino acid polypeptide by multiple gene duplications.

Three of the same repetitive regions are also found in the amino-terminal fragments of B (33, 34) and C2 (35)—i.e., the Ba and C2b fragments, respectively. B and C2 are also examples of C3b and C4b binding proteins, respectively, which contribute catalytic domains to the C3 and C5 convertases (58). A DIAGON analysis of murine protein H with factor B is shown in Fig. 3 *Middle*. Each of the three repeating elements of Ba has been shown to be encoded by separate exons (34, 57). In Fig. 3 *Bottom* is shown a comparison of murine protein H with human β_2-gpI, a protein of plasma origin previously shown to contain four repeating units. The function of β_2-gpI is unknown; however, its association with platelets and lipoprotein has been recognized. Two regions of the murine and the human interleukin 2 receptor are homologous to murine protein H (data not shown) (37, 38). More recently, the human gene structure for the interleukin 2 receptor has been reported, and limited homology with the Ba fragment of factor B has been noted (38). Specifically, sequences that are homologous with the repeating units of fragment Ba and protein H are encoded by exons 2 and 4 in humans as well as in mice. Deletion of exon 4 resulted in a dysfunctional molecule.

The significance of the repetitive regions in the above proteins is unknown at the present time. While each of these complement family members uniquely interact with C3b and/or C4b, no specific binding sequence has been identified. Further studies will be required to determine whether one or possibly several of these units constitute a binding domain. In human protein H, the C3b binding site has been determined to reside in the amino-terminal M_r 38,000 tryptic fragment corresponding to repeating units I through V in the mouse protein H sequence (16).

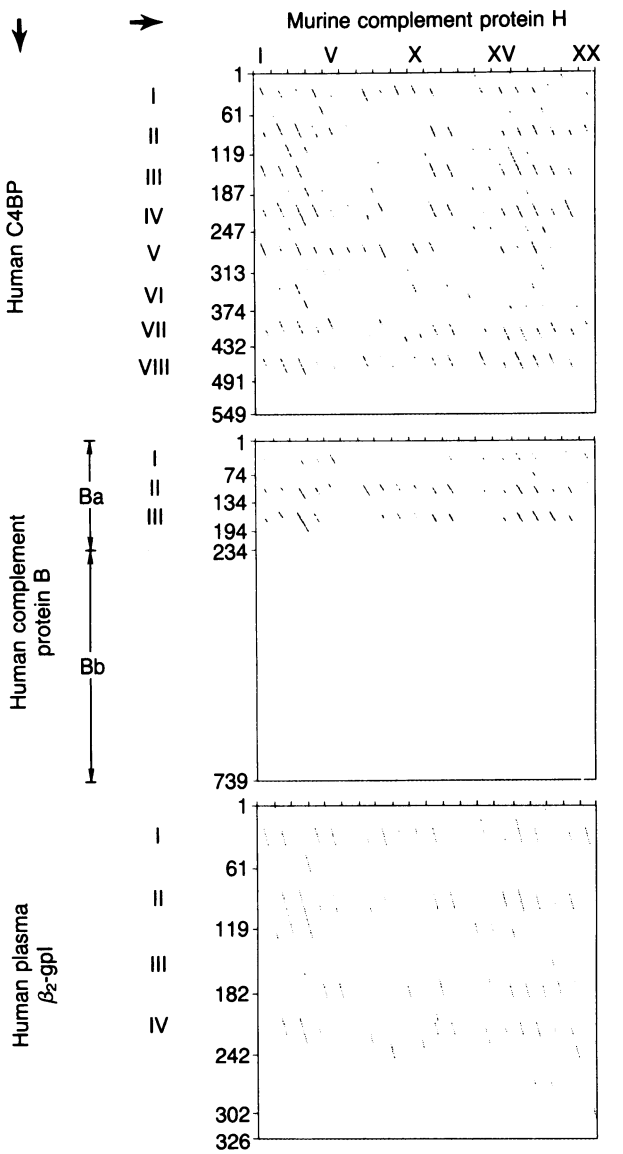

FIG. 3. DIAGON diagrams of the derived amino acid sequence for murine protein H versus C4BP (22) (*Top*), factor B (33, 34) (*Middle*), and β_2-gpI (36) (*Bottom*). The regions of homology are indicated by roman numerals, and their boundaries in the murine protein H sequence are as listed in Fig. 2. A percentage score of 285 and a span length of 25 was used. The division into repetitive regions is done as explained in the legend to Fig. 2.

We thank Dr. R. T. Ogata for making the mouse liver cDNA library available to us. Also, we thank Drs. R. A. Wetsel, N. P. H. Møller, R. T. Ogata, and D. J. Noonan for many helpful discussions and Mrs. Joan Celeste for typing the manuscript. Nomenclature in this paper follows that of the World Health Organization (30). These studies were supported by United States Public Health Service Grants AI 19222, AI 22214, and AI 17354. This is publication number 4213 IMM from the Department of Immunology, Research Institute of Scripps Clinic.

1. Kinoshita, T. & Nussenzweig, V. (1984) *J. Immunol. Methods* **71,** 247–257.
2. Kaidoh, T., Fujita, T., Takata, Y., Natsuume-Sakai, S. & Takahashi, M. (1984) *Complement* **1,** 44–51.
3. Natsuume-Sakai, S., Sudoh, K., Kaidoh, T., Hayakawa, J.-I. & Takahashi, M. (1985) *J. Immunol.* **134,** 2600–2606.
4. Horstmann, R. D. & Muller-Eberhard, H. J. (1985) *J. Immunol.* **134,** 1094–1100.
5. Horstmann, R. D., Pangburn, M. K. & Muller-Eberhard, H. J. (1985) *J. Immunol.* **134,** 1101–1104.
6. Whaley, K. & Ruddy, S. (1976) *Science* **193,** 1011–1013.
7. Whaley, K. & Ruddy, S. (1976) *J. Exp. Med.* **144,** 1147–1163.
8. Weiler, J. M., Daha, M. R., Austen, K. F. & Fearon, D. T. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 3260–3272.
9. Fearon, D. T. & Austen, K. F. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 1683–1687.
10. Fearon, D. T. & Austen, K. F. (1977) *J. Exp. Med.* **146,** 22–33.
11. Pangburn, M. K. & Muller-Eberhard, H. J. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 2416–2420.
12. Kazatchkine, M. D., Fearon, D. T. & Austen, K. F. (1979) *J. Immunol.* **122,** 75–81.
13. Sim, R. B. & DiScipio, R. G. (1982) *Biochem. J.* **205,** 285–293.
14. Jouvin, M.-H., Kazatchkine, M. D., Cahour, A. & Bernard, N. (1984) *J. Immunol.* **133,** 3250–3254.
15. Alsenz, J., Lambris, J. D., Schulz, T. F. & Dierich, M. P. (1984) *Biochem. J.* **224,** 389–398.
16. Alsenz, J., Schulz, T. F., Lambris, J. D., Sim, R. B. & Dierich, M. P. (1985) *Biochem. J.* **232,** 841–850.
17. DiScipio, R. G. & Hugli, T. E. (1982) *Biochim. Biophys. Acta* **709,** 58–64.
18. Fujita, T., Gigli, I. & Nussenzweig, V. (1978) *J. Exp. Med.* **148,** 1044–1051.
19. Gigli, I., Fujita, T. & Nussenzweig, V. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 6596–6600.
20. Ferreira, A., Takahashi, M. & Nussenzweig, V. (1977) *J. Exp. Med.* **146,** 1001–1018.
21. Kaidoh, T., Natsuume-Sakai, S. & Takahashi, M. (1981) *J. Immunol.* **126,** 463–467.
22. Chung, L. P., Bentley, D. R. & Reid, K. B. M. (1985) *Biochem. J.* **230,** 133–141.
23. Fearon, D. T. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 5867–6876.
24. Fearon, D. T. (1979) *J. Exp. Med.* **152,** 20–30.
25. Medof, M. E., Iida, K., Mold, C. & Nussenzweig, V. (1982) *J. Exp. Med.* **156,** 1739–1754.
26. Medof, M. E. & Nussenzweig, V. (1984) *J. Exp. Med.* **159,** 1669–1685.
27. Kinoshita, T., Lavoie, S. & Nussenzweig, V. (1985) *J. Immunol.* **134,** 2564–2570.
28. De Cordoba, S. R. & Rubinstein, P. (1984) *J. Immunol.* **132,** 1906–1908.
29. De Cordoba, S. R., Lublin, D. M., Rubinstein, P. & Atkinson, J. P. (1985) *J. Exp. Med.* **161,** 1189–1195.
30. World Health Organization Report (1981) *Bull. W.H.O.* **59,** 489.
31. Klickstein, L. B., Wong, W. W., Smith, J. A., Morton, C., Fearon, D. T. & Weis, J. H. (1985) *Complement* **2** (1), 44–45.
32. *Genetic Maps 1984*, ed. O'Brien, S. J. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), Vol. 3.
33. Mole, J. E., Anderson, J. K., Davison, E. A. & Woods, D. E. (1984) *J. Biol. Chem.* **259,** 3407–3412.
34. Morley, B. J. & Campbell, R. D. (1984) *EMBO J.* **3,** 153–157.
35. Bentley, D. R. & Campbell, R. D. (1985) *Complement* **2**(1), 9 (abstr.).
36. Lozier, J., Takahashi, N. & Putnam, F. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3640–3644.
37. Shimuzu, A., Kondo, S., Takeda, S.-i., Yodoi, J., Ishida, N., Sabe, H., Osawa, H., Diamantstein, T., Nikaido, T. & Honjo, T. (1985) *Nucleic Acids Res.* **13,** 1505–1516.
38. Leonard, W. J., Depper, J. M., Kanehisa, M., Kronke, M., Peffer, N. J., Svetlik, P. B., Sullivan, M. & Greene, W. C. (1985) *Science* **230,** 633–639.
39. Kristensen, T., Wetsel, R. A. & Tack, B. F. (1986) *J. Immunol.*, in press.
40. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113,** 237–251.
41. Okayama, H. & Berg, P. (1982) *Mol. Cell. Biol.* **2,** 161–170.
42. Okayama, H. & Berg, P. (1983) *Mol. Cell. Biol.* **3,** 280–289.
43. Sepich, D. S., Noonan, D. J. & Ogata, R. T. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 5895–5899.
44. Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7,** 1513–1523.
45. Grunstein, M. & Hogness, D. S. (1974) *Proc. Natl. Acad. Sci. USA* **72,** 3961–3965.
46. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
47. Bankier, A. T. & Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry*, ed. Flavell, R. A. (Elsevier/North-Holland Scientific, Limerick, Ireland), Vol. 85–08, pp. 1–34.
48. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
49. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3963–3965.
50. Staden, R. (1980) *Nucleic Acids Res.* **8,** 3673–3694.
51. Staden, R. (1982) *Nucleic Acids Res.* **10,** 4731–4751.
52. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
53. Bause, E. (1983) *Biochem. J.* **209,** 331–336.
54. Thibodeau, S. N., Palmiter, R. D. & Walsh, K. A. (1978) *J. Biol. Chem.* **253,** 9018–9023.
55. Steiner, D. F., Quinn, P. S., Chan, S. F., Marsh, J. & Tager, H. H. (1980) *Ann. N.Y. Acad. Sci.* **343,** 1–16.
56. Smith, C. A., Pangburn, M. K., Vogel, C.-W. & Muller-Eberhard, H. J. (1983) *Immunobiology* **164,** 298 (abstr.).
57. Campbell, R. D., Bentley, D. R. & Morley, B. J. (1984) *Philos. Trans. R. Soc. London Ser. B* **306,** 367–378.
58. Reid, K. B. M. & Porter, R. R. (1981) *Annu. Rev. Biochem.* **50,** 433–464.