Neural Cell Adhesion Molecule: Structure, Immunoglobulin-Like Domains, Cell Surface Modulation, and Alternative RNA Splicing

BRUCE A. CUNNINGHAM, JOHN J. HEMPERLY, BEN A. MURRAY, ELLEN A. PREDIGER, ROBERT BRACKENBURY, GERALD M. EDELMAN

The neural cell adhesion molecule, N-CAM, appears on early embryonic cells and is important in the formation of cell collectives and their boundaries at sites of morphogenesis. Later in development it is found on various differentiated tissues and is a major CAM mediating adhesion among neurons and between neurons and muscle. To provide a molecular basis for understanding N-CAM function, the complete amino acid sequences of the three major polypeptides of N-CAM and most of the noncoding sequences of their messenger RNA's were determined from the analysis of complementary DNA clones and were verified by amino acid sequences of selected CNBr fragments and proteolytic fragments. The extracellular region of each N-CAM polypeptide includes five contiguous segments that are homologous in sequence to each other and to members of the immunoglobulin superfamily, suggesting that interactions among immunoglobulin-like domains form the basis for N-CAM homophilic binding. Although different in their membrane-associated and cytoplasmic domains, the amino acid sequences of the three polypeptides appear to be identical throughout this extracellular region (682 amino acids) where the binding site is located. Variations in N-CAM activity thus do not occur by changes in the amino acid sequence that alter the specificity of binding. Instead, regulation is achieved by cell surface modulation events that alter N-CAM affinity, prevalence, mobility, and distribution on the surface. A major mechanism for modulation is alternative RNA splicing resulting in N-CAM's with different cytoplasmic domains that differentially interact with the cell membrane. Such regulatory mechanisms may link N-CAM binding function with other primary cellular processes during the embryonic development of pattern.

CELL ADHESION HAS LONG BEEN RECOGNIZED AS A MAJOR primary process in embryonic development. Only in the past decade, however, have some of the molecules responsible for cell-cell interactions been identified and characterized (*1, 2*). Analyses of the distributions and properties of these cell adhesion molecules (CAM's) have altered earlier notions that very large numbers of molecules with different specificities at the individual cell level are responsible for cell recognition in morphogenesis and histogenesis (*3*). Instead, it appears that a smaller number of CAM's of different binding specificities leads to border formation in early cell collectives and that repeated dynamic regulation of the expression and properties of the various CAM's at the surface of cells in such collectives is coordinated with movement, division, and cytodifferentiation events to yield tissue patterns. The role of CAM's in such processes has been shown at various sites (*4*), perhaps most dramatically in the processes of feather morphogenesis in which perturbation of CAM interactions led to marked changes in the periodic pattern of feathers in the skin as well as in the structure of feather precursors (*5*).

The neural cell adhesion molecule, N-CAM, is the best characterized CAM, and at least five forms of cell surface modulation regulate its expression and activity (*1*). N-CAM binding is homophilic (N-CAM on one cell binds to N-CAM on another) and small changes in its surface density lead to large changes in binding rates (*6*). The amount of N-CAM expressed (prevalence modulation), its asymmetric distribution on the cell surface (polarity modulation), and its expression relative to other adhesion molecules (differential prevalence modulation) all may influence patterning during development. The structural results described below emphasize the significance of two additional types of modulation: posttranslational modifications of the molecule (chemical modulation) and differential splicing of messenger RNA's (mRNA's) encoded by the single N-CAM gene leading to N-CAM molecules either with different modes of attachment to the cell membrane or different cytoplasmic domains.

N-CAM undergoes a number of posttranslational modifications that could modulate its expression and activity including addition of asparagine-linked oligosaccharides (*7*), phosphorylation of serine and threonine residues in the cytoplasmic domains (*8–10*), fatty acid acylation (*11*), and sulfation of asparagine-linked oligosaccharides (*8*). Outstanding among these modifications, however, are developmentally regulated and functionally significant alterations in the amount and distribution of α-2,8-linked polysialic acid (*12–15*), an unusual carbohydrate in a protein of vertebrate origin. The polysialic acid is attached to asparagine-linked oligosaccharides in the central portion of the molecule, outside the amino terminal binding region (*7, 16*). Although it does not participate directly in homophilic binding, this negatively charged carbohydrate is present in largest amounts in N-CAM from early embryonic cells and decreases with age in a tissue-dependent manner with a concomitant increase in N-CAM binding (*6, 17*).

The three prominent polypeptides of chicken N-CAM (Fig. 1A) appear to be identical from their amino termini to the region where they are associated with the membrane (*7, 9, 18, 19*). The two largest polypeptides are integral membrane proteins with substantial cytoplasmic domains. The larger of these (ld, large cytoplasmic domain polypeptide) contains 261 additional amino acids (*20*) in its

The authors are all members of the Laboratory of Developmental and Molecular Biology at The Rockefeller University, New York, NY 10021.

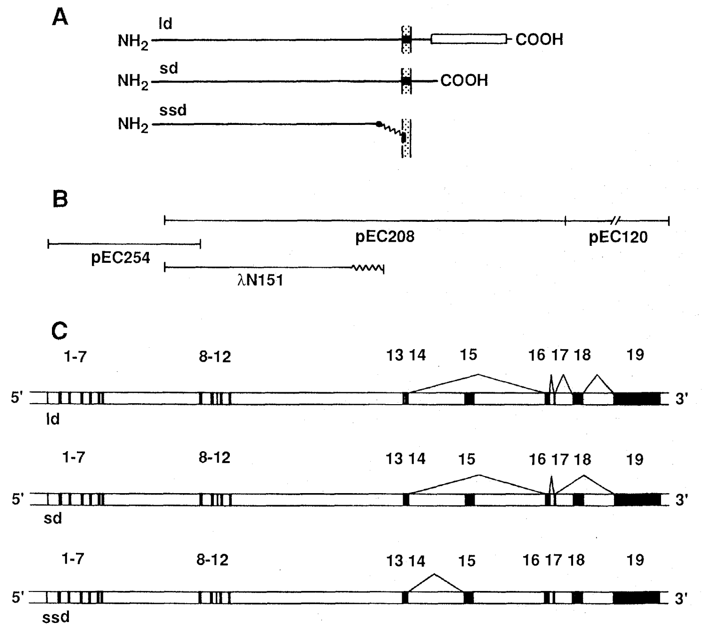

Fig. 1. Diagram of N-CAM polypeptides (A), cDNA clones (B) and the N-CAM gene (C). (**A**) The large cytoplasmic domain (ld) polypeptide contains an insert (open bar) not present in the small cytoplasmic domain (sd) polypeptide (*28*) although both contain identical membrane (stippling) spanning segments (closed bar), some common cytoplasmic sequences and identical carboxyl termini (*18*). The small surface domain (ssd) polypeptide lacks the membrane-spanning region and is attached to the membranes via a linkage (zigzag line) sensitive to phosphatidylinositol specific phospholipase C (*21, 23*) [adapted from figure 5 in (*21*)]. (**B**) Clones pEC208 and λN151 include the regions that distinguish the polypeptides from each other, whereas clones pEC254 and pEC120 complete the 5′ end of the coding sequence for the region common to all three polypeptides and extend the 3′ sequence through a large untranslated region common to the ld and sd mRNA's. The region of λN151 which differs from pEC208 is indicated by a wavy line. (**C**) The N-CAM gene in the chicken contains at least 19 exons (dark bars) spliced in various combinations (thin lines) (*25*). Exon 1 includes the amino terminus of all three polypeptides. Exons 1 to 14 are shared by all three polypeptides (ld, sd, and ssd). Exons 16, 17, and 19 appear in the mRNA's for the ld and sd chain, but not in the ssd chain. Exon 15 occurs only in the ssd mRNA and exon 18 is only in the ld mRNA [adapted from figures 2 and 5, in (*25*)]. (C) is not drawn to the same scale as (A) and (B).

cytoplasmic domain that are not present in the smaller (sd, small cytoplasmic domain polypeptide). The smallest polypeptide (ssd, small surface domain polypeptide) lacks the membrane spanning region of the ld and sd chains; analysis of complementary DNA's (cDNA's) reveals a different hydrophobic segment and no cytoplasmic domain for this chain (*21*). The ssd chain thus resembles other cell surface molecules that are attached to membranes by phosphatidylinositol-containing anchors (22). In accord with this hypothesis, the ssd chain of N-CAM, but not the ld or sd chain, is released from the cell membrane by phosphatidylinositol-specific phospholipase C (*21, 23*).

All of these N-CAM polypeptides are derived by alternative mRNA splicing from a single gene (*24, 25*), which is on chromosome 9 in mice (*26*) and on chromosome 11 in humans (*27*). The coding region of the chicken N-CAM gene contains at least 19 exons (*25*) spanning more than 50 kilobases (kb) (Fig. 1C). Fourteen exons are common to the three major polypeptides, and the individual chains thus arise by alternative splicing of the remaining five. Additional noncoding exons are located 5′ to exon 1 and will probably extend the size of the N-CAM gene to more than 70 kb. At least four N-CAM mRNA's have been detected. One (6.8 to 7.2 kb) codes for the ld chain (*28*), another (6.2 kb) specifies the sd chain (*20*), and the remaining two (4.2 and 6.0 kb) are detected by probes presumed to be specific for the ssd form (*21*).

```
CTGCAGGGGGGGGGGGGGGGGGGGGGGGCGGTGCTGCGGCGCTGTCAGTCAGCGGCGGCGGCGCCGGAGGCGGCCGGGGCGCAGCGGGCGGAGCGGGGACACGCGCGGAGCGGAGCGGA   120
PstI                                                                                   HaeIII

GCGGAGCGGCGCGAAGGGGACGCGGGGCAAAAACGGGAAGATCCATCCGGAGGGCCGCGTGGTGCTCCGCCCGCACGGAACGGCTATTTCGGAGCGGCCCCTCTCGCCCCCCCGCCGGCT   240
               ↳ pEC265                                 HaeIII                                        HaeIII

SIG                                                           NH2
-19                       -10                                 1                            10                            20
 M  L  P  A  A  A  L  P  W  T  L  F  F  L  G  A  A  A  S  L  Q  V  D  I  V  P  S  Q  G  E  I  S  V  G  E  S  K  F  F
GCGATGCTGCCCGCCGCCGCGCTGCCCTGGACGTTGTTTTTCTCGGAGCCGCAGCCTCTCTACAAGTGGATATTGTTCCAAGTCAGGGAGAGATCAGCGTTGGAGAATCCAAGTTCTTC   360

                          30                            40                            50                            60
 L  C  Q  V  A  G  E  A  K  Y  K  D  I  S  W  F  S  P  N  G  E  K  L  T  P  N  Q  Q  R  I  S  V  V  R  N  D  D  F  S  S
TTATGTCAAGTGGCAGGGGAAGCCAAATACAAAGACATTTCCTGGTTTTCCCCTAACGGTGAGAAGCTGACGCCCAACCAACAGCGCATCTCAGTGGTGCGAAACGATGACTTCTCCTCC   480

                          70                            80                            90                           100
 T  L  T  I  Y  N  A  N  I  D  D  A  G  I  Y  K  C  V  V  S  S  V  E  E  G  D  S  E  A  T  V  N  V  K  I  F  Q  K  L  M
ACCCTCACCATCTACAACGCCAACATTGATGATGCTGGCATCTATAAATGTGTTGTCAGCAGCGTGGAGGAGGGAGACTCTGAAGCCACCGTCAATGTGAAAATTTTCCAAAAGCTCATG   600

                         110                           120                           130                           140
 F  K  N  A  P  T  P  Q  E  F  K  E  G  D  D  A  V  I  Y  C  D  V  V  S  S  L  P  P  T  I  I  W  K  H  K  G  R  D  V  I
TTCAAGAATGCCCCCACTCCTCAGGAATTCAAGGAAGGGGATGATGCTGTGATTGTGTGTGATGTGGTCAGCTCGCTGCCTCCTACCATCATCTGGAAACACAAAGGCAGGGATGTCATC   720
                        EcoRI
             pEC265 ←

                         150                           160                  168
 L  K  K  D  V  R  F  I  V  L  S  N  N  Y  L  Q [ I  R  G  I  K  K  T  D  E  G  T  Y
CTAAAAAAAGATGTTCGGTTTATAGTGCTGTCCAACAACTACCTGCAG[ATCCGGGGAATCAAGAAAACAGATGAAGGGACGTAC(C)n CTGCAG ]
                                           PstI                                            PstI
```

Fig. 2. Nucleotide sequence of pEC254. The cDNA synthesis was primed with an oligonucleotide complementary to a sequence (boxed) near the 5′ end of pEC208, dG residues were added, and the cDNA was inserted into oligo(dC)-tailed pBR322 (*24, 29*). The resulting library yielded a number of N-CAM clones, the largest of which was pEC254. The nucleotide sequence of the large Pst I–Pst I fragment (nucleotides 1 to 768) of the cDNA insert of pEC254 was determined by dideoxynucleotide (*71*) and chemical degradation (*72*) sequencing techniques as previously described (*18*). Eco RI (GAATTC), Pst I (CTGCAG), and Hae III (GGCC) sites are indicated. The sequence determined from nucleotides 625 to 768 (shaded box) is identical to that at the 5′ end of pEC208. The sequence of the region enclosed in brackets (nucleotides 769 to 3′ end) was not determined directly, but is based on the sequence of pEC208 and the methodology used in the construction. The sequence of pEC265, which was obtained from a bacteriophage isolated from the λgt11 library constructed from the same cDNA, begins at a synthetic Eco RI linker attached at nucleotide 136 and terminates at the Eco RI site at nucleotide 625. The amino terminus (NH_2) of the N-CAM polypeptides is designated amino acid 1; the presumed signal sequence (SIG) begins at amino acid −19. Solid arrow underlines indicate amino acid sequences determined by protein chemical analysis of the intact polypeptide and of a CNBr fragment that began at residue 101.

We now report the complete primary structure of chicken N-CAM as deduced from cDNA sequences. We relate this structure to N-CAM binding and to the modulation of its expression and activity by changes in polysialic acid as well as in alternative RNA splicing. The complete structural analysis further corroborates the homology between N-CAM and members of the immunoglobulin superfamily, particularly the myelin-associated glycoprotein.

Sequence of cDNA Clones

The sequence of the N-CAM cDNA clone (pEC208), which extends from a natural Eco RI site within the N-CAM coding sequence through the carboxyl terminus of the ld and sd chains and into the 3′ untranslated regions of the mRNA for these two chains (Fig. 1B), has been reported (*18*). More recently, the sequence of a cDNA clone (λN151) that may encode the ssd chain was described (*21*).

The DNA sequences have now been extended in the 5′ direction to include sequences encoding the amino terminus of the polypeptides, a presumed hydrophobic signal sequence, and a noncoding region. The sequence of a cDNA clone corresponding to the large 3′ untranslated regions of the mRNA's for the ld and sd chains has also been determined. These results provide the complete amino acid sequences of the three major polypeptides of N-CAM and most of the nucleotide sequence of their mRNA's.

To obtain cDNA clones corresponding to the 5′ regions of the N-CAM mRNA's, we used an oligonucleotide complementary to a 17-bp sequence located 164 bp from the 5′ end of pEC208 to prime the synthesis of cDNA for libraries in both pBR322 (*24, 29*) and λgt11 (*30, 31*). These libraries were screened with a probe derived from the 5′ end of pEC208. The largest clone (pEC254, obtained in pBR322) was sequenced from its 5′ end to nucleotide 768; this region included an overlap of 144 nucleotides with the 5′ end of pEC208 (Fig. 2). Clone pEC254 codes for 187 amino acids (Figs. 1 and 2) that include the known amino terminal sequence of all three polypeptides (*7, 19*). This sequence is preceded by an apparent signal sequence that begins at an ATG initiation codon; the proposed site of cleavage of the signal sequence is identical to that predicted by the empirical rules of von Heijne (*32*). Preceding this sequence is an additional ~215 bp of apparently noncoding DNA sequence. An Eco RI fragment of the largest cDNA clone obtained from the λgt11 library was also subcloned (pEC265) and sequenced. It extends from nucleotide 136 of pEC254 (Fig. 2) to the natural Eco RI site at nucleotide 625.

Neither pEC208 nor pEC254 contained the sequences of our

Fig. 3. Nucleotide sequence of pEC120 (residues 1 to 2878), which contains most of the 3′ untranslated region of the ld and sd mRNA's. Two fragments of pEC120 produced by cleavage with Hind III at residue 1448 were subcloned into single-stranded M13 bacteriophage in both orientations. A series of smaller fragments generated from these clones by the deletion method of Dale *et al.* (*73*) was sequenced by the dideoxy method (*71*). The relation between pEC120 and pEC208 was determined by sequence analysis of a Taq I–Taq I fragment, derived from genomic clone λCN7 (*25*), which spanned the single natural Eco RI site at residue 1 (boxed). The locations of pEC001 and pEC020 (*24*) within pEC120 are indicated by arrows. Comparison with sequences represented in the GenBank database reveal two repetitive sequence elements in pEC120, an ATT repeat (dotted line A) and a region (dotted line B) that is strongly homologous to the CR1 chicken repetitive sequence. Two potential poly(A) consensus sites (boxed) are located at residues 804 and 2835. The arrow at residue 2869 indicates the point at which the cDNA sequence diverges from the genomic sequence (*25*), implying that this is the polyadenylation site.

original N-CAM cDNA clones pEC001 and pEC020 (*24*). Screening λgt11 libraries by hybridization with pEC001 and pEC020 gave a clone (pEC120) (Fig. 1B) that included both the pEC001 and pEC020 sequences but none of the pEC208 sequence (Fig. 3). The 2878-bp sequence of pEC120 terminated at a natural Eco RI site at its 5′ end and had no open reading frame, suggesting that it corresponds to a 3′ untranslated region throughout its length. To determine the relation between pEC208 and pEC120, a Taq I subfragment that hybridized with both pEC208 and pEC120 was isolated from an N-CAM genomic clone, λCN7 (*25*). Sequence analyses of this fragment showed that it contained a single Eco RI site. The 92 bp extending 5′ from this site were identical in sequence to the 3′ end of pEC208, and the sequence extending 3′ from this site was identical to the presumed 5′ end of pEC120 for at least 170 bp, confirming that the pEC208 and pEC120 sequences are adjacent in the N-CAM mRNA.

Near the 3′ end of pEC120 there is an AATAAA consensus polyadenylation signal that is followed, after 23 bases, by a stretch of 15 A residues. Only the first five of these A residues were found in the sequence of the genomic clone (*25*) corresponding to this region, suggesting that the remaining residues were added posttranscriptionally. Another potential poly(A) addition sequence was located approximately 2 kb upstream from the 3′ end of pEC120 (Fig. 3); there is as yet, however, no evidence to suggest that this site is used.

The plasmid pEC120 contains a region (residues 898 to 979) that is 79 percent identical to a portion of the chicken middle repetitive sequence element CR1 (*33, 34*). CR1 flanks several chicken gene complexes, the two copies being present in opposite orientations; it has been suggested (*34*) that CR1 may play a role in determining the availability of chromatin for transcription. The CR1-homologous region in pEC120, however, is in the opposite orientation to that observed in the 3′ end of other genes and may thus serve another function. Also, pEC120 contains many short stretches of redundant sequences. For example, residues 332 to 365 consist of repeated ATT residues, a structural feature that has been observed in the mouse gene for glial fibrillary acidic protein (*35*) and also flanking one clone of the mouse PR1 middle repetitive sequence (*36*). The significance of these similarities is unknown.

Amino Acid Sequence

The complete amino acid sequences of the ld, sd, and ssd chains of N-CAM as deduced from the nucleotide sequences of λN151, pEC208, and pEC254 are summarized in Fig. 4. The reading frame and portions of the sequence of pEC208 were previously verified by amino acid sequence analyses of CNBr and proteolytic fragments of the molecule (*18*). Additional fragments from this region and from portions of the molecule represented in pEC254 have now been sequenced by protein chemical techniques, and these data are indicated by solid underlines in Fig. 4. In all cases there is excellent agreement between the protein sequence and the cDNA sequence. We have detected only aspartic acid at position 4 and not the

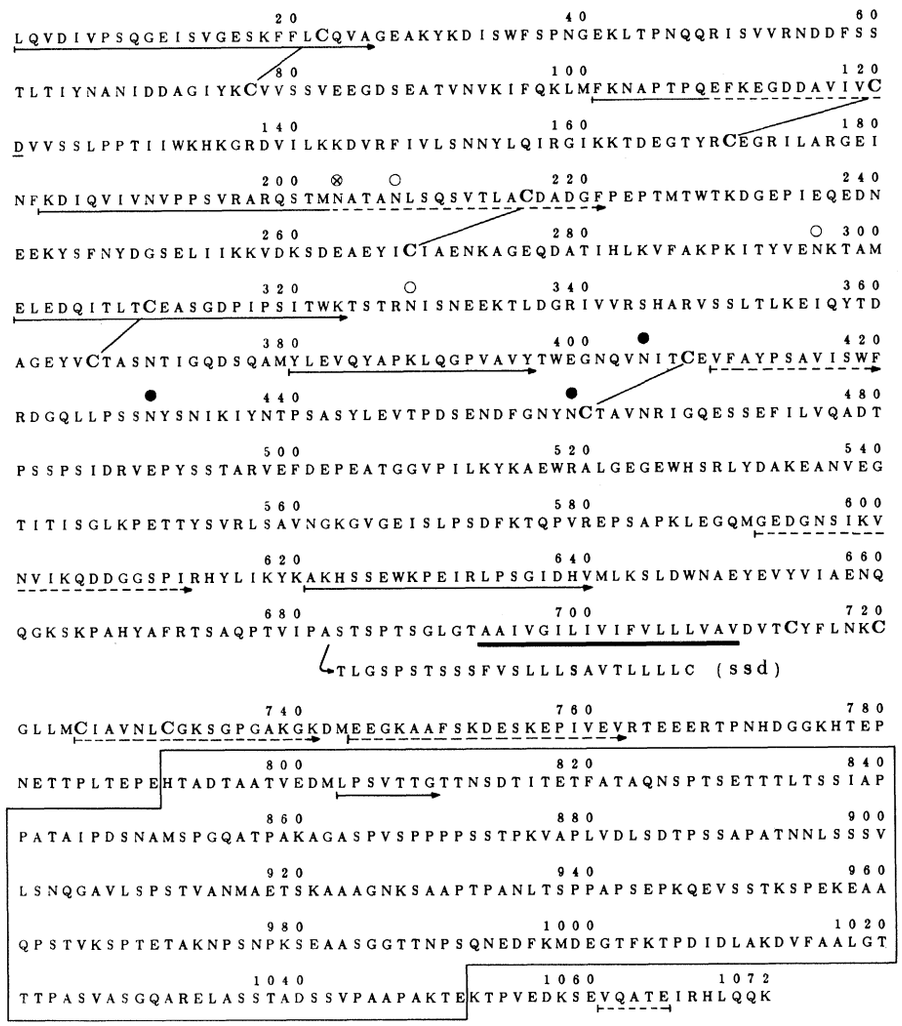

Fig. 4. Complete amino acid sequence of N-CAM [single letter code (*74*)]. The hydrophobic, presumptive membrane-spanning segment is indicated by a bar. The ld chain-specific residues are boxed (*20*). The residues coded for only by the ssd mRNA are shown on a separate line and include the carboxyl terminus of this chain (*21*); it is not yet clear whether some or all of these are removed during attachment of the presumed phosphatidylinositol-containing anchor (*21, 22*). Cysteines are in larger type and presumptive disulfide bonds connecting them are indicated. Potential attachment sites for asparagine-linked oligosaccharides are denoted by open circles and those that could carry the polysialic acid are indicated by closed circles. Dashed underlining denotes previously reported protein sequence data (*18*); solid underlines indicate new protein data. Chemical analysis indicates that the asparagine indicated by the cross within a circle (residue 203) probably contains covalently bound carbohydrate but the asparagine at residue 207 probably does not.

mixture of aspartic acid and glutamic acid reported for bovine N-CAM (*19*).

The cDNA sequence defines polypeptides of 115,467 daltons (ld), 89,625 daltons (sd), and 78,273 daltons (ssd), considerably smaller than the apparent molecular sizes of 160 kD, 130 kD and 110 kD, respectively, as measured on SDS-PAGE (polyacrylamide gel electrophoresis) for the polypeptides free of asparagine-linked oligosaccharides (*7, 13*). Nevertheless data from several sources indicate that the sequences account for the entire polypeptides. Clones pEC254 and pEC208 overlap, and each has been sequenced numerous times in both directions with no apparent gaps. Independent clones comparable to pEC208 and pEC254 have been isolated, and clones containing the 5′ sequences were obtained from both the λgt11 and pBR322 libraries. All amino acid sequences that have been obtained from fragments of N-CAM are included within pEC254 and pEC208. The size of the mRNA predicted for the ld chain (7 kb) is in excellent agreement with the size of the largest N-CAM mRNA (6.8 to 7.2 kb) and the size of the sd mRNA deduced from the cDNA's (6.2 kb) agrees with the observed value of 6.4 kb. The λN151-specific mRNA's (4.2 and 6 kb) (*21*) are larger than predicted from the cDNA sequence (3.0 kb), however, suggesting that additional sequences (most likely in the 3′ untranslated region) remain to be isolated for these mRNA's.

Further evidence that the deduced sequences account for the complete amino acid sequences was obtained by translation of mRNA synthesized in vitro from DNA constructions containing the coding sequences for the ld chain (Fig. 5) (*37*). The largest major translation product migrated with a molecular size of 150,000 daltons, comparable to the corresponding carbohydrate-free ld polypeptide and considerably larger than the size predicted (115,467 daltons) from the sequence of the cDNA. One-dimensional peptide mapping (Fig. 5, lanes 5 and 6) verified that this component contained N-CAM sequences. Additional components similar to those observed previously after in vitro translation of native mRNA (*24*) also were immunoprecipitated by antibodies to N-CAM. Calculations from the amino acid sequence suggested that the smaller components could have arisen by initiation at internal methionine residues. Similar constructions with clones encoding the ssd chain (predicted size 78,272 daltons) yielded a major polypeptide of 103,000 daltons. Other proteins, for example, c-myc (*38*), integrin (*39*), thrombospondin (*40*), and the LDL receptor (*41*), whose apparent molecular size on SDS-PAGE is larger than that predicted from the cDNA sequences have been reported, although the extent of difference varies. Such differences may reflect unusual features of specific sequences or more likely they may reflect as yet undefined variables in estimates of molecular weight by SDS-PAGE.

Potential N-CAM Binding Regions and Oligosaccharide Attachment Sites

N-CAM binding activity has been localized to a fragment (Frl, apparent molecular size 65 kD) that contains the amino terminus but lacks the bulk of the sialic acid (*7*); it thus should not extend beyond residue 400. A monoclonal antibody (anti–N-CAM 1) that detects determinants in Frl and blocks cell adhesion (*42*) was used to purify a chymotryptic peptide (>5 kD) that began at residue 183 and extended at least to residue 208. A second monoclonal antibody (anti–N-CAM 11) that blocks cell adhesion detects a fragment (23 kD) smaller than Frl. This smaller fragment includes the amino terminus of N-CAM, but does not react with monoclonal anti–N-CAM 1; it also includes the portion of N-CAM that binds heparin and presumably heparan sulfate (*43, 44*). These results suggest that two regions contained within the first three disulfide loops are involved in N-CAM binding. Monoclonal antibodies affecting N-CAM binding at two distinct sites in Frl have also been reported by others with (*43*) and without (*45*) supporting sequence data.

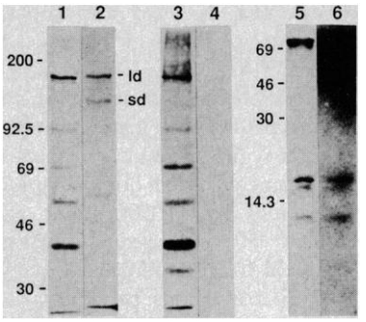

Fig. 5. In vitro expression of the N-CAM ld polypeptide. Constructions containing full-length ld coding sequences were transcribed and translated in vitro (*37*). Total translation products (lane 1) and translation products that were precipitated with rabbit antibodies to chicken N-CAM (lane 3) or nonimmune rabbit IgG (lane 4) were subjected to electrophoresis in the same 7.5 percent polyacrylamide gel, treated with Enhance (New England Nuclear), and exposed to x-ray film. For comparison, [^{35}S]methionine-labeled N-CAM was immunoprecipitated from tunicamycin-treated 10-day embryonic chicken brain cells with the monoclonal antibody anti–N-CAM 1 (*7*) and run on the same gel (lane 2); the ld and sd chains are indicated. Bands corresponding to the ld chains of lanes 1 and 2 were excised from similar gels and subjected to partial proteolytic digestion (*75*) with 0.5 μg of *Staphylococcus aureus* V8 protease; the digestion products were separated on a 15 percent polyacrylamide gel (*7*) and were detected by autoradiography. Lanes 5 and 6 correspond to digests of material equivalent to the ld bands in lanes 1 and 2, respectively. Migration positions of molecular size markers (kD) are indicated at the left of lanes 1 (for lanes 1 to 4) and 5 (for lanes 5 to 6).

The CNBr fragment that contains most, if not all, of the N-CAM polysialic acid (*16*) was isolated by repeated gel filtration on Sephacryl S300 in 0.1*M* ammonium bicarbonate. The partial amino acid sequence of this fragment was identical to that of N-CAM residues 380 to 396 (Fig. 4), supporting the previous conclusion (*18*) that the polysialic acid is on one or more of the three asparagine-linked oligosaccharides at positions 404, 430, and 459. Other evidence suggests that not all of the other four potential sites (residues 203, 207, 296, and 328) are glycosylated. Repeated sequence analysis of residues 203 to 220, for example, gave little asparagine at position 203, but asparagine was obtained in good yields at position 207, suggesting that oligosaccharide is attached at position 203, but not 207.

Internal Homology and Homology with Other Proteins

N-CAM shares many features with immunoglobulins (Ig's). The sequence of pEC208 revealed four contiguous segments of about 100 amino acids, all homologous to each other and to Ig domains (*18*). The complete sequence reported here reveals that there is a fifth such region (Figs. 4 and 6). Moreover, as in Ig domains, each of the five regions in the complete sequence of N-CAM contains a pair of cysteines that are spaced 50 to 56 amino acids apart. Because there are no interchain disulfide bonds or free SH groups in N-CAM (*42*), each pair of cysteines in the homology regions probably forms an intrachain disulfide loop (*18*) as they do in Ig domains. In support of this proposal, gel filtration and SDS-PAGE of individual Ig-like domains of N-CAM (obtained by CNBr cleavage) indicated that the domains were not linked to each other by disulfide bonds, so that the two cysteines within each homology region should be disulfide-bonded to each other. Another characteristic of Ig domains (*46, 47*) and closely related molecules such as β_2-microglobulin (*48, 49*) is extensive β structure. Alternating hydrophobic and hydrophilic residues usually found in such structures are seen in the areas around the conserved residues in the N-CAM sequence (Fig. 6). Preliminary circular dichroism studies of the N-CAM protein

indicate that N-CAM also contains β structure in the region represented by Fr1, which includes four of the five Ig-like homology regions (*50*).

Several observations suggest that the N-CAM gene diverged from the precursor of Ig genes before the divergence of variable and constant regions from each other. The N-CAM homology regions have characteristics of both variable and constant regions: the size of the regions (about 100 amino acids) and the distance between cysteines (50 to 56 residues) are comparable to Ig constant regions, but the N-CAM amino acid sequences more closely resemble those of Ig variable regions, particularly in the characteristic D-X-A(G)-X-Y-X-C sequence around the second cysteine (*51, 52*). There are seven residues that are conserved in all five N-CAM segments (Fig. 6) that also are highly conserved in Ig variable regions (*51, 52*). In addition, the aligned arginine and lysine residues at position 49 may correspond to those that are conserved at similar positions in Ig variable regions (*52*). The notion that the N-CAM precursor diverged early in evolution from the Ig precursor is also supported by the exon structure of the N-CAM gene (Figs. 1 and 6). Each of the five homology regions in N-CAM is specified by two exons (*25*), in contrast to the homology regions of previously known members of the Ig superfamily, which are specified by a single exon. This observation is consistent with previous hypotheses (*53, 54*) that Ig's may have evolved from a precursor gene smaller than that needed to encode an entire domain. Recently, it has been shown that the lymphocyte T4 protein, which is homologous to the Ig superfamily, is also specified by two exons (*55, 56*). Both N-CAM and the T4 protein would thus appear to resemble more closely the Ig precursor than other known members of the Ig superfamily.

Among the members of the Ig superfamily, the myelin-associated glycoprotein [MAG (*57–59*)] and Thy 1 antigen (*60*) are noteworthy because, like N-CAM, they are found in the nervous system. The sequences of N-CAM and Thy 1 are no more closely related than are N-CAM and other members of the extended Ig superfamily. However, MAG is more similar to N-CAM in both overall structure and sequence. Like N-CAM it has five consecutive regions of about 100 amino acids each with potential disulfide loops. Three of these regions are homologous to each other and to N-CAM and, of these, two are more similar to N-CAM (25 and 29 percent identity) (*57, 58*) than N-CAM is to other members of the Ig superfamily (15 to 23 percent identity). Moreover, there appear to be two forms of MAG with differing cytoplasmic domains that result from alternative RNA splicing (*58*). These results are consistent with proposed roles for MAG (*57–59*) and (less persuasively) for Thy 1 (*52, 60*) as mediators of cell adhesion.

Other nonimmunoglobulin-like molecules have some weak similarities to N-CAM. In nearly all of these cases, however, the similarity extends for only a few residues. For example, the GPIIb protein of the platelet aggregation system and the apparently related LFA-1 and Mac-1 molecules from leukocytes have modest similarity with the first 15 to 20 residues of N-CAM (*61*), but are also similar to a portion of the internal segment in the ld chain (residues 811 to 820). The significance of such limited similarity to both external and intracellular segments of N-CAM is unclear. Limited similarity to fibronectin (*18*) also has been detected, but this too spans only about seven residues in the center of N-CAM (residues 543 to 549). Also, N-CAM does not contain R-G-D (*62*) or R-E-D-V sequences (*63*) involved in the cell-binding activity of fibronectin.

Evolutionary and Functional Considerations

The complete amino acid sequences of the ld, sd, and ssd polypeptides (Fig. 4) provide a firmer basis for describing N-CAM function. Outstanding among the structural features is the similarity of a large portion of the molecule with members of the Ig superfamily, many of which have recognition and binding functions. By analogy with these molecules, the results presented here suggest

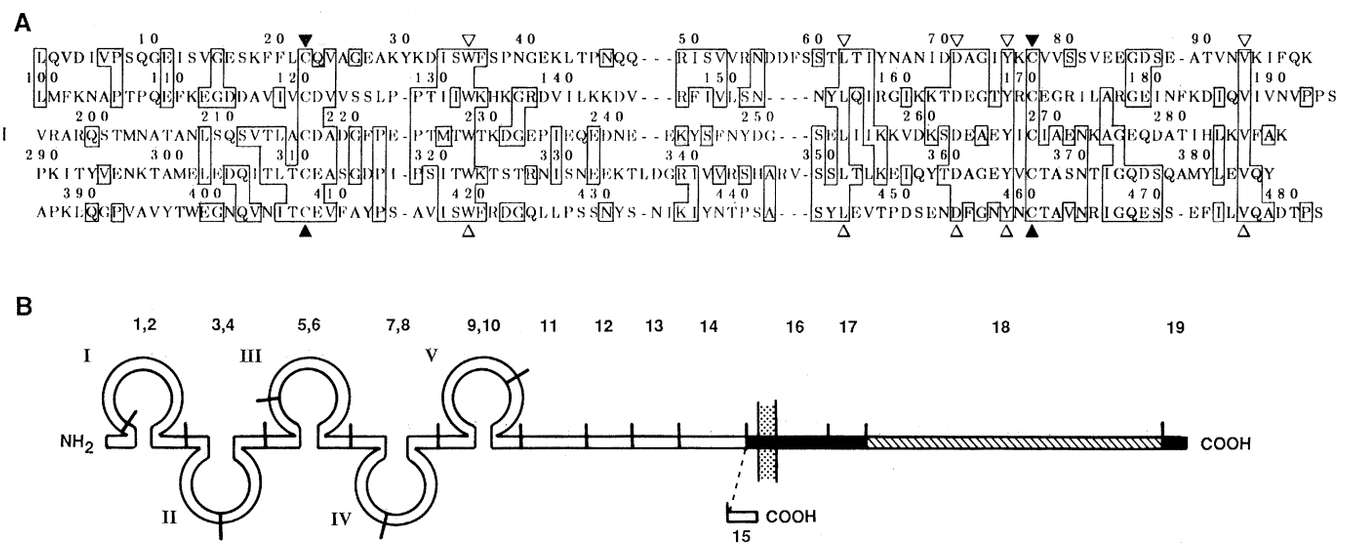

Fig. 6. (**A**) Alignment of the five internally homologous segments (I to V) common to the ld, sd, and ssd chains, showing their similarity with each other and with members of the Ig superfamily. Residues are numbered consecutively from the amino terminus of the mature N-CAM polypeptides. Sequences were aligned pairwise initially according to the program FASTP (*76*), and finally by inspection to give the greatest overall match (*18*). Residues identical in all five N-CAM regions and highly conserved among Ig-like proteins are marked with triangles; the cysteines proposed to be involved in intradomain disulfide bonds are indicated by closed triangles. Residues identical in two or more sequences are boxed. (**B**) Model of N-CAM showing homologous loops (I to V) in the regions common to all three polypeptides (open bar). The base of each loop corresponds to the proposed intradomain disulfide bond. The cell membrane is indicated by stippling and the extracellular and intracellular regions are to the left and right, respectively. The membrane-spanning region and carboxyl terminal segment common to the ld and sd chain are indicated by the solid bar, while the cytoplasmic domain unique to the ld chain is indicated by the hatched bar. The unique segment of the ssd chain is indicated by the open bar and dashed line below. Numbers 1 to 19 correspond to exons in the chicken N-CAM gene (*25*) and their relative boundaries in the protein are noted by transverse lines [adapted from figure 5 in (*25*)].

that the amino terminal portion of N-CAM is folded into five relatively compact domains (Fig. 6) that serve its binding activity and related functions. It is attractive to consider two related ideas: that the specialized functions of Ig superfamily molecules were derived from a precursor involved in cell adhesion and that domain-to-domain interactions involving the homology regions are a major structural basis of homophilic (*6*) N-CAM binding. Such domains could act directly by pairing with domains of N-CAM molecules on apposing cells or could first interact pairwise with other N-CAM molecules on the same cell to generate the necessary sites for cell-cell interaction.

From previous studies (*7*), we deduce that the direct binding regions should be among domains I to IV (Fig. 6). Because homophilic binding (binding between identical molecules from apposing cells) requires complementarity, we anticipate that at least two such regions are involved. As indicated above, studies with monoclonal antibodies are in accord with this hypothesis, which is also consistent with the strong evolutionary conservation of N-CAM binding function (*64*). One of the epitopes recognized by a monoclonal antibody (anti–N-CAM 1) that inhibits cell adhesion is near the disulfide loop of domain III; the epitope for another inhibitory monoclonal antibody (anti–N-CAM 11) is amino terminal to this. Because of the relatively large size of the antibody Fab' fragments, however, such studies obviously cannot determine the binding regions with precision.

Unlike the Ig's, there appears to be no variation in the amino acid sequences of the binding region of N-CAM that could lead to differences in binding specificity. No evidence for extensive polymorphism was apparent in either the cDNA or amino acid sequences, although some single base differences were noted between the overlapping sequences of λN151 and pEC208 (*21*); the few differences between the cDNA and amino acid sequences were usually at positions where the amino acid sequence data were inconclusive. As we have emphasized here, an alternative to the idea of differentiation in binding specificities is that the activity of N-CAM is regulated by the cells it ligates via a series of cell surface modulation events (*1*), apparently triggered by local signals produced by developing cell collectives linked by CAM's (*4, 5*). A strong example is provided in the fifth homology region which carries the polysialic acid that modulates the kinetics of N-CAM binding. By charge repulsion and steric hindrance, the long polymers of negatively charged sialic acid could act cis to alter conformation and accessibility of the binding region or alter trans binding to N-CAM on apposing cells. The sulfation of N-CAM oligosaccharides (*8*) could also add negative charges to this same region.

Whereas the extracellular portions of the three N-CAM polypeptides are the same, the membrane-associated segments and internal domains differ as the result of alternative mRNA splicing. The ld and sd polypeptides are integral membrane proteins that are mobile in the membrane (*65, 66*) and can be acylated with fatty acids (*11*). The ssd chain does not span the membrane, however, but is attached by a phospholipid anchor (*20, 22*) and as such is probably even more mobile at the cell surface. The cytoplasmic domains of the sd and ld chains allow them to interact with molecules in the cell cortex and to be further modulated by phosphorylation of serine and threonine residues. The larger cytoplasmic domain of the ld chain provides even more opportunity for cytoplasmic interactions and it also contains additional phosphorylation sites. Beyond these carboxyl terminal sequences, the large 3′ untranslated regions of the mRNA's for the ld and sd polypeptides may play an important role in regulating their expression.

Although the detailed relationships of these structural differences to N-CAM expression and activity in morphogenesis are still unknown, useful clues may come from pursuing the observation that the polypeptides are expressed at different times and places during development. As a result of alternative RNA splicing, the ld chain first appears after neurulation and only on neural tissues (*20, 67*) and the ssd chain is expressed in significant amounts only later at the time of glial maturation (*21, 68*). In addition to the potential role of differential splicing in N-CAM release from cells, it presumably influences the mobility, relative prevalence, and polarity of N-CAM on various cell surfaces as a function of time and cellular position and could thus have profound effects on embryonic development. The different cytoplasmic domains might affect cortical events involving second messengers or might possess enzymatic activity. A related question is whether this form of N-CAM modulation serves to link N-CAM binding to other cellular events such as cell differentiation, division, growth, and movement. If so, it would be a central feature of the global modulation (*69*) of cell surfaces that can accompany such cellular events during the development of pattern.

Note added in proof: Barthels *et al.* (*70*) have reported the sequence analysis of cDNA clones corresponding to the mouse ssd polypeptide. This polypeptide is identical in length and 85 percent identical in amino acid sequence to the chicken ssd sequence reported here.

REFERENCES AND NOTES

1. G. M. Edelman, *Annu. Rev. Cell Biol.* **2**, 81 (1986).
2. For general reviews of families of adhesion molecules, see the papers from several laboratories in *The Cell in Contact: Adhesions and Junctions as Morphogenetic Determinants*, G. M. Edelman and J.-P. Thiery, Eds. (Wiley, New York, 1985).
3. R. W. Sperry, *Proc. Natl. Acad. Sci. U.S.A.* **50**, 703 (1963).
4. K. L. Crossin, C.-M. Chuong, G. M. Edelman, *ibid.* **82**, 6942 (1985).
5. W. J. Gallin, C.-M. Chuong, L. H. Finkel, G. M. Edelman, *ibid.* **83**, 8235 (1986).
6. S. Hoffman and G. M. Edelman, *ibid.* **80**, 5762 (1983).
7. B. A. Cunningham, S. Hoffman, U. Rutishauser, J. J. Hemperly, G. M. Edelman, *ibid.*, p. 3116.
8. B. C. Sorkin, S. Hoffman, G. M. Edelman, B. A. Cunningham, *Science* **225**, 1476 (1984).
9. G. Gennarini, G. Rougon, H. Deagostini-Bazin, M. Hirn, C. Goridis, *Eur. J. Biochem.* **142**, 57 (1984).
10. J. M. Lyles, D. Linnemann, E. Bock, *J. Cell Biol.* **99**, 2082 (1984).
11. B. C. Sorkin and B. A. Cunningham, unpublished results.
12. J. B. Rothbard, R. Brackenbury, B. A. Cunningham, G. M. Edelman, *J. Biol. Chem.* **257**, 11064 (1982).
13. G. Rougon, H. Deagostini-Bazin, M. Hirn, C. Goridis, *EMBO J.* **1**, 1239 (1982).
14. J. Finne, U. Finne, H. Deagostini-Bazin, C. Goridis, *Biochem. Biophys. Res. Commun.* **112**, 482 (1983).
15. C.-M. Chuong and G. M. Edelman, *J. Neurosci.* **4**, 2354 (1984).
16. K. L. Crossin, G. M. Edelman, B. A. Cunningham, *J. Cell Biol.* **99**, 1848 (1984).
17. R. Sadoul, M. Hirn, H. Deagostini-Bazin, G. Rougon, C. Goridis, *Nature (London)* **304**, 347 (1983).
18. J. J. Hemperly, B. A. Murray, G. M. Edelman, B. A. Cunningham, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3037 (1986).
19. G. Rougon and D. R. Marshak, *J. Biol. Chem.* **261**, 3396 (1985).
20. B. A. Murray *et al.*, *J. Cell Biol.* **103**, 1431 (1986).
21. J. J. Hemperly, G. M. Edelman, B. A. Cunningham, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9822 (1986).
22. M. G. Low, M. A. J. Ferguson, A. H. Futerman, I. Silman, *Trends Biochem. Sci.* **11**, 212 (1986).
23. H.-T. He, J. Barbet, J.-C. Chaix, C. Goridis, *EMBO J.* **5**, 2489 (1986).
24. B. A. Murray *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5584 (1984).
25. G. C. Owens, G. M. Edelman, B. A. Cunningham, *ibid.* **84**, 294 (1987).
26. P. D'Eustachio, G. C. Owens, G. M. Edelman, B. A. Cunningham, *ibid.* **82**, 7631 (1985).
27. C. Nguyen, M.-G. Mattei, J.-F. Mattei, M.-J. Santoni, C. Goridis, B. R. Jordan, *J. Cell Biol.* **102**, 711 (1986).
28. B. A. Murray, J. J. Hemperly, E. A. Prediger, G. M. Edelman, B. A. Cunningham, *ibid.*, p. 189.
29. T. Maniatis, E. G. Fritsch, J. Sambrook, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1982), p. 230.
30. W. J. Gallin, E. A. Prediger, G. M. Edelman, B. A. Cunningham, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2809 (1985).
31. R. A. Young and R. W. Davis, *Science* **222**, 778 (1983).
32. G. von Heijne, *Nucleic Acids Res.* **14**, 4683 (1986).
33. W. E. Stumph, C. P. Hodsson, M.-J. Tsai, B. W. O'Malley, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6667 (1983).
34. W. E. Stumph, M. Baez, W. G. Beattie, M.-J. Tsai, B. W. O'Malley, *Biochemistry* **22**, 306 (1983).
35. J. M. Balcarek and N. G. Cowan, *Nucleic Acids Res.* **13**, 5527 (1985).
36. R. Kominami *et al.*, *J. Mol. Biol.* **165**, 209 (1983).
37. The Hae III–Eco RI fragment from pEC254 (Fig. 2, nucleotides 217 to 630) was inserted into Sma I–Eco RI digested pGEM-1 (Promega Biotec). The resulting plasmid was linearized with Eco RI, treated with calf intestinal alkaline phospha-

tase, and the Eco RI fragment of pEC208 was inserted. Correctly oriented constructs were identified by restriction mapping of the resulting transformants with Hae III. Purified DNA was linearized by digestion with Sst I (which cuts once within the 3′ untranslated region of pEC208), and blunt ends were formed by treatment with the Klenow fragment of DNA polymerase I; this template was used for in vitro transcription by bacteriophage SP6 RNA polymerase [D. A. Melton, P. A. Krieg, M. R. Rebagliati, K. Zinn, M. R. Green, *Nucleic Acids Res.* **12**, 7035 (1984)]. The resulting RNA was translated in a rabbit reticulocyte lysate system (Promega Biotec) in the presence of [^{35}S]methionine (50 μCi) and immunoprecipitated with rabbit polyclonal antibodies to N-CAM (*23*).
38. M. R. Gazin, M. Rigolet, J.-P. Briand, M. H. V. Van Regenmortel, F. Galibert, *EMBO J.* **5**, 2241 (1986).
39. J. W. Tamkun *et al.*, *Cell* **46**, 271 (1986).
40. J. Lawler and R. O. Hynes, *J. Cell Biol.* **103**, 1635 (1986).
41. R. Yamamoto *et al.*, *Cell* **39**, 27 (1984).
42. S. Hoffman *et al.*, *J. Biol. Chem.* **257**, 7720 (1982).
43. G. J. Cole, A. Loewy, N. V. Cross, R. Akeson, L. Glaser, *J. Cell Biol.* **103**, 1739 (1986).
44. S. Hoffman and B. A. Cunningham, unpublished results.
45. M. Watanabe, A. L. Frelinger III, U. Rutishauser, *J. Cell Biol.* **103**, 1721 (1986).
46. G. M. Edelman, *Biochemistry* **9**, 3188 (1970).
47. L. M. Amzel and R. J. Poljak, *Annu. Rev. Biochem.* **48**, 961 (1979).
48. B. A. Cunningham, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35**, 1171 (1976).
49. J. W. Becker and G. N. Reeke, Jr., *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4225 (1985).
50. S. Hoffman, G. M. Edelman, B. A. Cunningham, unpublished results.
51. E. A. Kabat, T. T. Wu, H. Bilofsky, M. Reid-Miller, H. Perry, *Sequences of Proteins of Immunological Interest* (National Institutes of Health, Bethesda, MD, 1983).
52. A. F. Williams, A. N. Barclay, M. J. Clark, J. Gagnon, in *Gene Expression during Normal and Malignant Differentiation* (10th Sigrid Juselius Foundation Symposium), L. C. Andersson, C. G. Gahmberg, P. Ekblom, Eds. (Academic Press, New York, 1985), p. 125.
53. A. Bourgois, *Immunochemistry* **12**, 873 (1975).
54. A. D. McLachlan, *Protides Biol. Fluids Proc. Colloq.* **28**, 29 (1980).
55. T. Maddon and R. Axel, personal communication.
56. D. R. Littman and S. N. Gettner, *Nature (London)* **325**, 453 (1987).
57. M. Arquint *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 600 (1987).
58. J. L. Salzer, W. P. Holmes, D. R. Colman, *J. Cell Biol.*, **104**, 957 (1987).
59. J. G. Sutcliffe, R. J. Milner, T. M. Shinnick, F. E. Bloom, *Cell* **33**, 671 (1983).
60. A. F. Williams and J. Gagnon, *Science* **216**, 696 (1982).
61. I. F. Charo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8351 (1986).
62. E. Ruoslahti and M. D. Pierschbacher, *Cell* **44**, 517 (1986).
63. M. J. Humphries, S. K. Akiyama, M. Komoriya, K. Olden, K. M. Yamada, *J. Cell Biol.* **103**, 2637 (1986).
64. S. Hoffman, C.-M. Chuong, G. M. Edelman, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 6881 (1984).
65. W. E. Gall and G. M. Edelman, *Science* **213**, 903 (1981).
66. E. G. Pollerberg, M. Schachner, J. Davoust, *Nature (London)* **324**, 462 (1986).
67. E. G. Pollerberg, R. Sadoul, C. Goridis, M. Schachner, *J. Cell Biol.* **101**, 1921 (1985).
68. G. Gennarini *et al.*, *J. Neurosci.* **6**, 1983 (1986).
69. G. M. Edelman, *Science* **192**, 218 (1976).
70. M. Barthels *et al.*, *EMBO J.* **6**, 907 (1987).
71. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
72. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1983).
73. R. M. K. Dale, B. A. McClure, J. P. Houchins, *Plasmid* **13**, 31 (1985).
74. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
75. D. W. Cleveland, S. G. Fischer, M. W. Kirschner, U. K. Laemmli, *J. Biol. Chem.* **252**, 1102 (1977).
76. D. J. Lipman and W. R. Pearson, *Science* **227**, 1435 (1985).
77. We thank V. Malik, E. Pichersky, and J. MacGregor for help with aspects of DNA sequencing; M. Burgoon for assistance with purification of the anti–N-CAM 1–reactive peptide; J. Salzer, D. Colman, J. Roder, and R. Axel, who communicated results prior to publication; and P. Ferrie, S. Lieberman, V. Sone, L. Tomchak, M. Petruzziello, B. Weerbrouck, and J. Diller for technical assistance. Supported by NIH grants HD-16550, HD-09635, AM-04256, and by a Senator Jacob Javits Center for Excellence in Neuroscience award, NS-22789, and a grant from Johnson and Johnson. Protein sequence analysis was carried out by the Rockefeller University Sequencing Facility, supported in part by the U.S. Army Research Office for the purchase of equipment.

Research Articles

Cloning of Large Segments of Exogenous DNA into Yeast by Means of Artificial Chromosome Vectors

DAVID T. BURKE, GEORGES F. CARLE, MAYNARD V. OLSON

Fragments of exogenous DNA that range in size up to several hundred kilobase pairs have been cloned into yeast by ligating them to vector sequences that allow their propagation as linear artificial chromosomes. Individual clones of yeast and human DNA that have been analyzed by pulsed-field gel electrophoresis appear to represent faithful replicas of the source DNA. The efficiency with which clones can be generated is high enough to allow the construction of comprehensive libraries from the genomes of higher organisms. By offering a tenfold increase in the size of the DNA molecules that can be cloned into a microbial host, this system addresses a major gap in existing experimental methods for analyzing complex DNA sources.

STANDARD RECOMBINANT DNA TECHNIQUES INVOLVE THE in vitro construction of small plasmid and viral chromosomes that can be transformed into host cells and clonally propagated. These cloning systems, whose capacities for exogenous DNA range up to 50 kilobase pairs (kb), are well suited to the analysis and manipulation of genes and small gene clusters from organisms in which the genetic information is tightly packed. It is increasingly apparent, however, that many of the functional genetic units in higher organisms span enormous tracts of DNA. For example, the bithorax locus in *Drosophila*, which participates in the regulation of the development of the fly's segmentation pattern, encompasses approximately 320 kb (*1*). The factor VIII gene in the human, which encodes the blood-clotting factor deficient in hemophilia A, spans at least 190 kb (*2*). Recent estimates of the size of the gene that is defective in Duchenne's muscular dystrophy suggest that this

The authors are in the Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110.