Retroviral integrase domains: DNA binding and the recognition of LTR sequences

Esther Khan⁺, Joseph P.G.Mack¹, Richard A.Katz, Joseph Kulkosky and Anna Marie Skalka*

Fox Chase Cancer Center, Institute for Cancer Research, Philadelphia, PA 19111 and

¹Crystallography Laboratory, NCI – Frederick Cancer Research and Development Center, PO Box B, Frederick, MD 21702, USA

Received October 18, 1990; Revised and Accepted December 14, 1990

ABSTRACT

Integration of retroviral DNA into the host chromosome requires a virus-encoded integrase (IN). IN recognizes, cuts and then joins specific viral DNA sequences (LTR ends) to essentially random sites in host DNA. We have used computer-assisted protein alignments and mutagenesis in an attempt to localize these functions within the avian retroviral IN protein. A comparison of the deduced amino acid sequences for 80 retroviral/retrotransposon IN proteins reveals strong conservation of an HHCC N-terminal 'Zn finger'-like domain, and a central D(35)E region which exhibits striking similarities with sequences deduced for bacterial IS elements. We demonstrate that the HHCC region is not required for DNA binding, but contributes to specific recognition of viral LTRs in the cutting and joining reactions. Deletions which extend into the D(35)E region destroy the ability of IN to bind DNA. Thus, we propose that the D(35)E region may specify a DNA-binding/cutting domain that is conserved throughout evolution in enzymes with similar functions.

INTRODUCTION

Retroviruses are distinguished from other RNA viruses by two steps in their replication cycle: reverse transcription and integration of viral DNA (provirus) into the host chromosome (for a general review see [1]). These events are catalyzed by products of the retroviral *pol* gene: reverse transcriptase (RT) and the integration protein or integrase (IN), respectively (Fig. 1) [2–4]. During transcription of viral RNA into DNA by RT, long terminal repeats (LTRs) are formed at the ends of the linear DNA products (see Fig. 3A). Each LTR is comprised of sequences derived from the 5' and 3' ends of the viral RNA (denoted U5, R and U3). The linear viral DNA appears to be the precursor of the integrated provirus [5,6] and the ends of the LTRs contain *cis*-acting sequences that are required for integration [7–9]. There is no apparent sequence specificity for host cell DNA integration sites although a preference has been noted [10].

However, during the integration reaction, characteristic sequence alterations of both the viral DNA and the host DNA occur; two nucleotides are lost from each end of the viral DNA, and cellular DNA (4 to 6 base pairs) is duplicated at the integration site. The integration reaction of retroviruses is similar to the DNA breakage/joining reactions characteristic of prokaryotic transposable elements, in that both types of reactions include transposition from 'within' a larger DNA and duplication of host cell target sequences [11–14]. Both types of elements also encode proteins required for their own transposition.

In ASLV, whose prototype is the Rous sarcoma virus (RSV), IN is synthesized as part of the *gag-pol* precursor and is formed by N- and C-terminal proteolytic processing [15–17]. The ASLV IN domain is also contained in the C-terminal third of the β chain of RT (Fig. 1). ASLV IN possesses DNA binding, endonuclease, and DNA joining activities that can be demonstrated *in vitro* [13,18]. Purified ASLV IN can preferentially remove two T residues from the 3' ends of duplex oligonucleotides that correspond to the ends of the LTRs (5'.....CA↓TT3'), exposing the conserved CA dinucleotide that is joined to host DNA *in vivo* [19]. A similar endonuclease activity has been detected for IN protein of Moloney Murine Leukemia Virus (MoMLV) [20,21]. Under certain conditions, ASLV IN also has the potential to make a staggered 6 bp cut, which is consistent with formation of the 6 bp direct repeat found flanking the integrated DNA [22,23].

Analysis of the deduced amino acid sequences of IN proteins may reveal conserved structural elements associated with known IN functions: 1) for specific recognition and removal of 2 nucleotides from the ends of the viral DNA 2) for binding and cutting the host target DNA, and 3) joining host and viral DNA ends. Such considerations suggest that IN may contain two DNA recognition sites. The endonuclease and integrase activities of ASLV and MLV IN require Mg^{2+} or Mn^{2+} but do not require an exogenous source of energy such as ATP [13,14,18]. A mechanism which derives energy from the integration reaction itself, might involve a covalent bond between the DNA intermediates and IN [24], such as occurs in other strand

* To whom correspondence should be addressed

⁺ Present address: Center for Advanced Molecular Biology, University of the Punjab, New Campus, Lahore-20, Pakistan

exchange proteins such as topoisomerases [25]. Because two sterically and temporally coordinated reactions (one for each end of the viral DNA) are required for successful integration of the viral DNA, it is reasonable to suppose that IN functions as a dimer, and DNA footprinting studies with ASLV IN support this hypothesis [26]. IN proteins may, therefore, contain a region(s) involved in protein-protein interactions.

In this paper we compare the deduced amino acid sequence of ASLV IN with that of other retroviruses, retrotransposons and bacterial insertion elements. Computer-assisted alignments have revealed two highly conserved regions [denoted HHCC and D(35)E] that may represent DNA binding or cutting domains. (A preliminary report on this alignment has been presented [27].) Bacterially-produced IN mutants were used to determine the roles of these regions in DNA binding/cutting. Surprisingly, the N-terminal HHCC region, reminiscent of a transcription factor DNA binding domain [28], was neither necessary nor sufficient for binding. Rather, the D(35)E region near the middle of the protein may contribute to DNA binding.

MATERIALS AND METHODS

Enzymes and biochemicals. Calf intestinal phosphatase was from Boehringer Mannheim (Indianapolis, IN). *Bal31* exonuclease, T4 DNA ligase, Klenow fragment of *E. coli* DNA polymerase I, and restriction enzymes were from New England Biolabs (Beverly, MA) and Bethesda Research Labs (Bethesda, MD). Lithium diiodosalicylate was from Sigma (St. Louis, MO). The affinity purification of protein A fusion proteins was carried out using IgG Sepharose 6 fast flow, from Pharmacia (Piscataway, NJ). ³H-thymidine 5'-triphosphate, ¹²⁵I-labelled protein G, and α-[P³²]dXTPs were purchased from Amersham Corporation (Arlington Heights, IL). d(AT) copolymer was from Pharmacia.

Sequence analysis. Protein sequences were deduced from the DNA sequence or extracted directly from the GCG [29] and the Los Alamos HIV [30] databases. Sequence manipulations used programs associated with the GCG package [31]. Data used were those entered up to December 31, 1989. Initial searches used the program PROFILESEARCH, accepting proteins which has a sequence similar to that of the known sequences of IN from RSV, HIV and MoMLV and which were located in an appropriate part of the retroviral/retrotransposon genome (3' end of the *pol* gene or 5' end of the *TyB* gene). Wider searches were conducted using the program TFASTA. A high scoring protein was accepted as an IN only if it also showed the above six conserved residues. An earlier computer search of 5 closely related IN sequences [28], showed 16 conserved residues. On extending the analysis to 80 retroviral and retrotransposon IN proteins, we found that only 6 of these residues (HHCCDE) are totally conserved in the regions shown in Figures 2 and 3.

Residue numbering: IN numbering is according to the residue position in RSV IN.

Protein sequences. Sequences accepted are listed by abbreviation (organism name/virus name, accession name/synonyms). A selected member of each retroviral group was chosen for display in the alignment figures (2 and 3). When several members of a group have the same name, the sequence selected for display is marked by an (*):

Retroviral/Retrotransposon: Arabidopsis (Arabidopsis Tal-3, X13291); BaEV (Baboon endog. V, PCBBEVXX/PCBCG); BLV (bovine leukemia virus, GNLJGB(*), BLVCG/BLVGAGA); Drosophila 17.6 (DROTN176); Drosophila 297 (DROIS297); Drosophila 412 (DROIS412/GNFF42); Drosophila 1731 (Drosophila, DROTN1731); Drosophila Copia (Drosophila DROVLPHR/OFFFCP/DROTNCOPIA/DROCOPIA); Drosophila gypsy (DROGYPPF1A/GNNFFG1/DROGYPSY); EIAV (Equine infectious anemia virus GNLJEV(*), GNLJEV/EIAVCG); Human-D (HRDLTRA); MMTV (Murine mammary tumor virus, GNMVMM(*), GNMSIP/MUSIAPIL3, MMTENVGR, MMTPROCG); FeEV (Feline endogenous virus (FE6POLG); FIV (Feline IV FIV14); FLV (Feline leukemia virus, M19392); Hamster IAP (GNHYIH); HIV1 (GNVWVL(*), MNCG/HIVMNCG, HIVNL43/HIVNY5/HIVBR, HIVMAL, HIVELICG, HIVZ2Z6, HIVH3BH5, GNVWH3/HIVPV22/HIVBH102, HIVHXB2CG, GNVWLV/HIVBRUCG, GNVWA2/HIVSF2CG, HIVRF); HIV2 (HIV2ISY, GNLJG2/HIV2ROD, HIV2NIHZ); HTLV1 (HL1PRCAR(*), HL1PROP/GNLJGH); HTLVII (GNLJH2/HIVV2CG; HumERVKA (Human endogenous virus ERVKA, GNHUEH/HUMERVKA); HumER41 (Human endogenous virus 41, HUMER41); HSPuENV (Human spuma, HSPUENV); MLV (Murine leukemia virus, MLVPOLA/GNMSRE, MLVENVR/GNMVVRV, GNVWK (AKR MLV), GNMVGV/MLOCG (AKV MLV); MoMLV (Moloney murine leukemia virus, GNMV1M/MLMCG); Mouse IAP (GNMSIA(*), MUSFLIAP); MuEV (Mouse endogenous virus C3H/He MUSMULVEA); RSV (Rous sarcoma virus, Prague, ALRCG/GNFV1R); SRV1-D (Simian SRV1-D, SIVRV1CG/GNLJSA); SRV2-D (Simian SRV2-D, SIV2DCG); SMP-D (Simian Mason Pfizer-D, GNLJMP/SIVMPCG); SIV (Sooty mangabey SIVSMMH4, macaque SIVMM142, macaque GNLJG3/SIVMM251, African green monkey GNLJG5/STLV-IIIagm, GNLJG4/SIVAGMTYO); SNV (Spleen necrosis virus/avian reticuloendotheliosis virus/REV, FOVDAR); Squirrel monkey RV (PCSPOL); Tobacco (Tobacco retrotransposon X13777); Tyl-17 Yeast (YSCTY117/B23496); Tyl-H3 (yeast YSCTY1H3A); Ty-pY109 (yeast YSCTY109); Ty3-1 (yeast YSCTY31A); Ty3-2 (yeast M23367); Ty912 (yeast SYNYSCTE); Visna (Visna lentivirus, Icelandic strain LV11, GNLJVS/VLVCG).

Insertion sequences. B. pertussis RS1 (BPETERRA, M28220); E. coli IS2 (ECOINS2K); E. coli IS3 (INS3); E. coli IS3411 (TRN3411); P. syringae IS51 (PSEIS51A); L. casei ISL1 (LCAISL1); E. coli INS150CG (INS150CG); S. agalactiae IS861 (STR861IS); L. lactis IS904 (LACNISB); S. dysenteriae IS (SHDSHTA); S. sonnei IS600 (SHSIS600).

Sequence alignment. The functional N-terminal end of IN was located by alignment of the HHCC region with that of RSV IN. The C-terminal end could not be located in the general case. Multiple sequence alignment used the minimal mutation distance method of Hein [32], which also gives a phylogenetic tree for the alignment. This program can only align sequences of similar length. The comparison was therefore restricted to those residues which aligned with the first 180 residues of RSV IN.

Residue similarity. For Figures 2 and 3, residues were scored by vertical alignment into six small groups of amino acids with similar side chain properties: ST, ILMV, DENQ, FYW, GA, KR; or into two large groups: ILMVIFYW, which are generally

hydrophobic, and QNSTDE, which are hydrogen bonding or polar. The shading patterns used were: white on black—single type of residue present in at least 50% of the proteins, or if a member of a small group was present in at least 75% of the proteins, or if a member of a large group was present in at least 90% of the proteins; white on grey—member of a small group present in at least 50% of the proteins, or member of a large group present in at least 75% of the proteins; black on white—non-conserved positively charged residues. If the two conserved HH residues in the HHCC motif are vertically aligned by hand, other groups come into register. These, which are also shaded in Figure 2, include: the two conserved H groups in the HHCC motif, the hydrophobic residues near position 1 and 25, and the hydrophobic residues at position 16 which are an alternative alignment to the residues at position 19. The two halves of Figure 3 (IN proteins and insertion sequences) were scored separately. Residues from one of the protein groups which belonged to groups shaded in the other half, were shaded white on grey.

Secondary structure prediction. Secondary structure was predicted using PEPLOT [33], based on the Chou and Fasman method [34] and the Garnier prediction method [35] of the GCG package. Predictions were compared with the results of CD spectroscopy on RSV IN, which showed IN to have 17% α -helix, 32% β -sheet, 18% turns and 33% irregular structure [36].

Construction of carboxyl and amino terminal deletion clones and *E. coli* expression. Plasmid pRC23-p32 [37], linearized with *Xho*I which cleaves just downstream of the 3' end of the IN coding sequences, was used as substrate for *Bal*31 digestion to make carboxyl terminal deletions. The samples were then digested with *Eco*RI, which cleaves 4 bp upstream of the 5' end of IN [16]. The resultant *Eco*RI-*Bal*31 digested fragments were cloned in a second vector which introduced *Bam*HI sites at the 3'-ends of the IN segments. After repair of the IN 5' *Eco*RI sites, *Eco*RI/*Bam*HI fragments were cloned in the *Sma*I and *Bam*HI sites of the polylinker region of the protein A fusion vector, pRIT2T (Pharmacia, Piscataway, NJ), which contains the lambda phage P_R promoter. Ligation of the repaired *Eco*RI site at the 5' end of IN, with *Sma*I-digested pRIT2T resulted in an in-frame fusion of protein A and IN translational reading frames. This cloning strategy generated a set of IN C-terminal deletions which were fused in translational frame to the C-terminus of protein A. The complete IN fusion was made in a similar fashion. The N-terminal deletion mutant, IN(17–286), was made by digesting the plasmid pRIT2T-IN with *Eco*RI and *Bss*HII, removing the overhangs with mungbean nuclease, and religating. The DNAs were used to transform the *E. coli* strain MC1061 [38], which also expressed a temperature-sensitive repressor from a separate plasmid, pRK248cIts [39]. The presence of fusion proteins was verified by electrophoresis of proteins from induced cell lysates on polyacrylamide gels [40], followed by immunoblot analysis utilizing a goat anti-reverse transcriptase serum [16].

Oligonucleotide-directed mutagenesis. Oligonucleotides were used to introduce changes in specific amino acid residues in the HHCC domain of IN, by published procedures [41,42]. The changes were: His9 to Asn or His13 to Asn as single mutations in pRC23p32. Additional nucleotides were changed without altering the encoded amino acids in order to create new *Dra*I or *Ssp*I restrictions sites. All deletions and site-directed mutations

were confirmed by the sequencing method of Maxam and Gilbert [43]. The mutagenic oligonucleotides were:

5' AGAGAGGCTAAAGATTTAAATACCGCTCTCCATA: His13(CAT) to Asn(AAT)
*Dra*I

5' CTTTCATACCGCTCTCAATATTGGACCCCGCG: His9(CAT) to Asn(AAT)
*Ssp*I

Isolation and purification of fusion proteins. For the isolation of the Staphylococcus protein A fusions, overnight cultures were diluted 1:100 in 30 ml of M9 medium and grown at 30°C to early stationary phase (O.D.₆₀₀ = 0.7–0.9). Expression from the lambda P_R promoter in the pRIT2T derivatives was derepressed by a shift in temperature to 42°C for 2–3 hr. Cells were pelleted by centrifugation and suspended in 2 ml of ice-cold TST buffer (50 mM Tris-HCl, pH 7.6, 150 mM NaCl and 0.05% Tween). Cells were lysed by sonication and a soluble fraction was prepared by centrifugation in a Beckman 70 Ti rotor at 30,000 rpm for 30 min at 4°C. The supernatants were diluted with the addition of 3 ml cold TST buffer and incubated on ice for 10 min. The clear supernatants were then loaded on 1 ml IgG Sepharose 6FF columns which had been pre-equilibrated with 10 column volumes of TST. The protein A fusions were eluted with 0.3 M lithium diiodosalicylate (pH 8.0) after a 12 ml wash with TST. Twelve fractions of 250 μ l each were collected and 15 μ l aliquots analyzed on 10% SDS polyacrylamide gels. Fractions containing the IN fusion protein (detected by Coomassie staining) were then assayed for DNA binding and/or endonuclease activity. Active fractions were dialyzed against cold Buffer A (50 mM Tris-HCl pH 7.8, 2 mM DTT, 0.1 mM EDTA and 10% glycerol) using the microdialysis system (BRL) and stored in 40% glycerol at –20°C.

Purification of mature, non-fusion IN protein. Procedures for purification of IN from AMV particles have been described [19]. For purification of IN from bacteria, *E. coli* strain MC1061 containing the mature IN expression vector pRC23-p32, or H9N or H13N derivatives, and the plasmid encoding the temperature-sensitive λ CI repressor was grown in modified M9 medium at 30°C to an O.D.₆₀₀ of 0.5. Expression of the protein was induced by increasing the temperature to 42°C as described by Terry *et al.* [37]. Bacteria were harvested by centrifugation and pellets from 2 liters of culture were suspended in 20 ml of sonication buffer (5 M NaCl, 50 mM Tris-HCl pH 7.4, 10% glycerol, 4 mM DTT and 0.1 mM EDTA). The cells were lysed by sonication (35 sec $\times$ 20 pulses) and bacterial debris removed by centrifugation. Then 6% polyethylene glycol 8000 (PEG) and 4% Dextran 500 (w/w) were added to the cell lysate and the mixture was allowed to emulsify at 4°C for 2–3 hr. Phase separation was accelerated by low speed centrifugation at 4°C and the top phase was withdrawn. The protein phase was extensively dialyzed in P11 buffer (100 mM NaCl, 50 mM Tris-HCl pH 7.4, 4 mM mercaptoethanol, 0.1 mM EDTA, 10% glycerol) to remove the PEG and Dextran. The dialyzed sample was then loaded onto a 75 ml phosphocellulose column (15 $\times$ 2.5 cms) that had been pre-equilibrated with P11 buffer. The protein was eluted with a linear gradient of 0.1 to 1.2 M NaCl. The fractions containing the IN protein were pooled, dialyzed in storage buffer (40% glycerol, 100 mM NaCl, 50 mM Tris-HCl pH 7.4, 4 mM mercaptoethanol and 0.1 mM EDTA) and stored at –70°C. For final purification, protein solutions were diluted 4-fold, applied to a polyU Sepharose column, and

eluted as described previously [37]. At this stage, wild-type IN and H9N, H13N mutant proteins were >90% pure as judged by Coomassie staining after gel electrophoresis.

DNA binding assay. ^3H -d(AT) copolymer (500 cpm/pmol nucleotide), the substrate for the DNA binding assay was prepared as described by Terry [37]. Column fractions were assayed in a 100 μl reaction containing 50 ngs of protein sample, 75 pmoles substrate DNA, 10 mM Tris-HCl pH 8.0, 5 mM DTT, 10 mM KCl and 10 mM MgCl_2 . Reactions were incubated at 37°C for 3 min, then diluted in 2 ml dilution buffer (same as reaction

buffer except 40 mM KCl) and passed slowly through nitrocellulose filters. Filters were counted to measure the DNA retention [37].

Oligonucleotide nicking assay. Standard conditions described by Katzman [19] were used to assay 50–100 ng samples of purified fusion proteins or non-fused IN proteins. Incubation was at 37°C in 2 mM Mn^{2+} or Mg^{2+} for the times indicated in the figures. Reactions were analyzed on 20% polyacrylamide gels containing 8 M urea.

Fig. 1. Genetic origin of IN and map of the protein A-IN gene fusion. The top line shows a genetic map of RSV, the source of IN sequences used in these studies. The second line shows the sequences included in the *gag-pol* mRNA which serves for translation of the viral *gag-pol* precursor protein, Pr180 (heavy horizontal line), which is made by a frameshift at the site indicated by a break on the line. The α and β subunits of reverse transcriptase (RT) and integrase (IN) are produced by proteolytic processing of Pr180 at the time or after virus budding. The bottom map shows the operator and fused coding sequences used to construct the vector which expresses the protein A-IN fusion proteins. Amino acid positions are indicated showing HHCC and D(35)E regions.

RESULTS

Sequence alignment. The multiple sequence alignment used here [32] compares sequences on a residue-by-residue basis. It employs a modified version of the Dayhoff Mutation Data Matrix [44] to determine the probability that amino acid differences in proteins of related function are the result of mutations from a common ancestor. It produces an alignment in which sequences are arrayed such that those closest to one another can be derived by the minimum amount of mutation. Aligned residues are thus presumed to be related in an evolutionary and hence functional sense [45]. No other information, such as the tendency of a particular set of residues to form a secondary structural unit, is implicitly included in the alignment. Presumably it is possible for a series of mutations (insertions or deletions) to move a conserved structural element horizontally within each sequence. In such a case, the ability to infer the presence of a conserved structure depends on recognition of secondary structure-forming elements, such as the constellation of histidine and cysteine residues which are the hallmarks of a Zn finger.

Fig. 2. Alignment in the retroviral/retrotransposon HHCC region. Sequences are grouped by phylogenetic relatedness. Amino acids are designated in the single letter code. Only a representative subset of the 80 sequences compared is illustrated. For identification of the total set and description of shading criteria, see Materials and Methods.

Alignment of residues corresponding to the first 180 amino acids (a.a.) of RSV IN (286 a.a.), gave a phylogenetic tree consistent with a classification of retroviruses based on genome organization. Thus, the alignment reflects the evolutionary history of IN and should contain information about conserved structure. The sequences are ordered in Figure 2 and Figure 3 to show phylogenetic relatedness, adjacent pairs of sequences being most closely related. The first sequences of these figures (RSV to SNV) contain retroviruses and endogenous retroviruses, while the last sequences (Ty3-2 to Copia) contain retrotransposons and endogenous retroviruses. The alignment shows that the integrases vary in length between 280–450 residues, with an N-terminal region (RSV residues 9–40) that contains conserved histidines and cystidines (the HHCC region, Fig. 2) and a region of conserved length (for RSV, starting at residue 121) which has the sequence D(35)E (Fig. 3).

The HHCC residues, although present in all IN proteins, are not vertically aligned. Thus, any conserved structural element associated with these residues must have moved during evolution, to different positions in the sequence. Consistent with this hypothesis, there are hydrophobic residues at position 17 for the sequences MoMLV through Gypsy which may functionally align with the hydrophobic residues at position 19. The only IN in our

comparison that did not include an HHCC sequence was from the molecular clone HIV2NIHZ which has the sequence HHCV [46]. It may be relevant that the full length NIHZ DNA is not infectious for HUT-78 human T cells, (G. Franchini, personal communication). We examined the possibility of a conserved structure in the HHCC regions using secondary structure prediction programs. These showed that most HHCC sequences had either a helix breaking G or P halfway along the finger, or else a predicted turn which allows the sequence to fold back along itself. However, there was no unifying structural theme for the two legs on either side of the turn.

The D(35)E region encompasses about 50 residues; its central feature is a conserved D (RSV residue 121) followed by 35 residues and then a conserved E (RSV residue 157). (There are small variations in the 35 residue length for two INs from the distantly related *Drosophila* retrotransposons.) The region is highly conserved. At 16 positions, conservative replacements occur at least 75% of the time, while at another 16 positions conservative replacements occur at least 50% of the time. The conservation is sufficiently strong that the remaining non-conserved residues can be aligned without gaps or insertions (Fig. 3, top section). The secondary structure predictions are more consistent for this region than for the HHCC region. The

Fig. 3. Alignment of the D(35)E motif in retroviruses, retrotransposons, and bacterial insertion sequences. As in Figure 2 only a subset of sequences compared is illustrated.

sequence around the conserved D is usually TDNG or similar. The predicted structure is beta sheet up to a turn at the TDNG, then random coil or α helix for the first quarter of the motif. The region around E157 is also α helix. Similar D(35)E sequences of the same length and which aligned without insertions or deletions, were found in a number of bacterial insertion sequences (IS) (Fig. 3, bottom). The evolutionary relationship among some of these IS elements has already been noted [47]. The sequence similarity between the retroviruses and IS elements extends in both directions from the D and E residues, covering about 50 residues.

Construction and analysis of IN fusion proteins. In order to facilitate the biochemical analyses of the conserved regions in ASLV IN, we used fusion proteins produced in bacteria. In this way, a series of different mutant proteins could be purified using a common domain. The 5' end of the IN coding region was joined to the 3' end of the *S. aureus* protein A gene (Fig. 1) and the fusion protein was expressed in *E. coli*. Although the protein A-IN fusion was not visible as a strong band on a Coomassie brilliant blue stained gel, it could be detected by immunoblot analysis utilizing anti-IN antibodies [16] (data not shown). Such analysis indicated that more than 50% of the fusion protein was in the soluble fraction of the *E. coli* lysate.

Soluble fusion proteins were bound to IgG affinity columns via the protein A component and then eluted. The recovered proteins were greater than 70% homogeneous as judged by analysis on SDS PAGE gels. As shown in Figure 4, a slower migrating protein co-eluted with the protein A-IN fusion in the eluate fractions. This protein was not related to protein A or IN as determined by immunoblotting (not shown). A number of the faster migrating bands, however, were IN-related. We presume that they represent premature termination or degradation products. Some fusion protein-containing fractions that were eluted late from the columns contained a precipitate of unknown origin and were inactive in the assays described below.

Activities of IN fusion proteins. Prior to any attempts to analyze altered fusion proteins, it was necessary to determine whether

Fig. 4. Purification of protein A-IN by IgG affinity chromatography. *E. coli* cells induced for expression of the protein A-IN(wt) fusion protein were lysed by sonication as described in Materials and Methods. Cell debris was removed by centrifugation. The supernatant fraction was applied to an IgG-Sepharose column and, after washing, the fusion protein was eluted with 0.3 M lithium-diiodosalicylate. Fractions were dialyzed to remove salt and samples applied to SDS PAGE. The gel was stained with Coomassie brilliant blue. The arrow indicates the location of full-length protein A-IN(wt) protein.

the wild-type (wt) IN fusion exhibited the DNA binding and endonuclease activities characteristic of the non-fused viral protein. Sequence-independent DNA binding activity was tested using a nitrocellulose filter binding assay with ³H-labelled d(AT) copolymer as the substrate [37]. This activity may represent the ability of IN to select random host integration sites. Results of a comparison of viral IN and the protein A-IN(wt) fusion (Table I) show that the binding activity of the IN-fusion was slightly lower than that of the viral protein (on a molar basis). This was not surprising since the protein undergoes a rigorous column elution which may lead to partial denaturation. As a negative control, binding experiments were carried out using protein A that was expressed from the parent vector and purified in parallel. These protein A-containing fractions showed little binding activity above background (Table 1).

Endonuclease activity of the IN-fusion was tested using an oligonucleotide cleavage assay. A 15mer representing the minus strand of the RSV LTR U3 end was ³²P-labelled at its 5' terminus (with polynucleotide kinase) and annealed to a

Table I: Comparison of DNA binding activity of viral IN and the protein A-IN(wt) fusion

Protein	cpm	pMoles DNA bound (-Bgd)
None	1023	
viral IN	38762	39
protein A	3202	3
protein A-IN(wt)	45731	46

Protein A and the protein A-IN(wt) fusion were eluted from IgG columns, dialyzed and ca. 50 ngs were used for binding assays. The viral IN control reaction contained 10 ng of purified AMV IN. Proteins were incubated with 75 pMoles ³H-d(AT) copolymer as described in Materials and Methods. Binding was detected by retention on nitrocellulose filters.

Fig. 5. Oligonucleotide assay of Mn²⁺-dependent IN endonuclease activity. A. Conditions for preparation of the substrate and enzymatic digestion are described in Materials and Methods. Incubation with 100 ngs of protein was for 15 min. Specificity is indicated by release of two nucleotides producing a labeled oligonucleotide which is two nucleotides shorter (indicated by an arrow and -2) as detected in this sequencing gel. B. The U3 substrate consists of duplex 15mers representing the U3 terminus of the upstream LTR with the radiolabel (³²P) on the 5' end of the minus strand deoxynucleotide. Arrowheads denote the known *in vivo* sites of cleavage.

complementary 15mer (Fig. 5B). The duplex was incubated with the fusion protein and samples of the reaction were analyzed on 20% sequencing gels. Specificity was indicated by the removal of two T residues from the 3' end, producing a labelled oligonucleotide product that is two nucleotides shorter (−2) than the substrate [19]. The endonuclease activity of the protein A-IN(wt) fusion was significantly lower than the viral protein in the presence of Mg^{2+} (data not shown). However, using Mn^{2+} as the divalent cofactor, the activities of these proteins were similar (Fig. 5A). It may be relevant that when IN is fused to other *pol* sequences, as it is in the beta subunit of RT (see Fig. 1), endonuclease activity is also observed only in the presence of Mn^{2+} [48,22]. The major endonucleolytic product of the protein A-IN(wt) fusion is an oligonucleotide two residues shorter than the substrate (indicating the loss of two T residues). In this assay, the fusion protein exhibited even greater specificity for the −2 site than the authentic viral protein. Accordingly, in subsequent analyses of the mutant fusion proteins we used Mn^{2+} as the cofactor. As expected, protein A expressed and purified in parallel showed no specific cleavage of LTR substrates. The small amount of −1 product observed with all bacterially-produced proteins was assumed to reflect contaminating bacterial exonucleases.

DNA binding activity of mutant IN proteins. Having established that the protein A-IN(wt) fusion displayed activities similar to the viral protein, we introduced deletions and single amino acid changes into the IN coding sequences (Fig. 6). The mutations included C-terminal truncations [denoted protein A-IN(1–63), -IN(1–105), -IN(1–149), -IN(1–206), and -IN(1–252), to indicate the IN-derived N-terminal amino acids retained] and an N-terminal truncation [(denoted protein A-IN(17–286) to indicate the IN-derived C-terminal amino acids retained)]. The amino acid changes were the relatively conservative histidine to asparagine substitutions at positions 9 (H9N) or 13 (H13N). The full-length wild-type and the mutant protein A-IN fusions were analyzed on

SDS PAGE gels to normalize the amount of protein used in both DNA binding and subsequent endonuclease activity assays.

The relative d(AT) copolymer DNA binding activity of these mutant proteins is summarized in Table II. The results show that deletion of sequences from the C-terminal third of the protein A-IN fusions [e.g. IN(1–252) and IN(1–206)] caused *ca.* 60% decrease in d(AT)copolymer binding. C-terminal deletions which extended into the conserved D(35)E region of the fusion proteins [e.g. IN(1–149)], or further [IN(1–105), IN(1–63) or IN(1–19)], showed no detectable binding. In contrast, the N-terminal deletion which removed both histidine residues of the HHCC region [IN(17–286)], or fusion proteins with substitutions predicted to disrupt a Zn finger structure (H9N or H13N), showed slightly increased binding of the copolymer relative to the wild-type IN fusion protein control. We conclude from these results that sequences in the N-terminal HHCC region do not contribute to this sequence-independent DNA binding activity. Indeed, in this construct they seem to partially mask the capacity of the protein to bind the copolymer. On the other hand, sequences in or near to the D(35)E region appear to be necessary for binding and sequences near the C-terminus may enhance this binding.

In experiments summarized in Table III, we investigated the capacity of fusion proteins with mutations in the HHCC region

Table II: Relative DNA binding activity of protein A-IN mutants

Protein	Percent d(AT)bound
protein A-IN (wt)	100
protein A-IN(1–252)	65,27
protein A-IN(1–206)	39
protein A-IN(1–149)	0*
protein A-IN(1–63)	0
protein A-IN(1–19)	0
protein A-IN(H9N)	120,180
protein A-IN(H13N)	133,218
protein A-IN(17–286)	156,247

* Binding was equal to or less than background with no protein or protein A alone. Conditions were as described in Table I. The table shows results from several independent experiments each of which included protein A, protein A-IN(wt) and no protein controls. The range of d(AT)copolymer bound by the protein A-IN(wt) in these experiments was 13–30 pmoles. For each experiment, values were normalized to this positive control. Duplicate numbers show results from separate experiments.

Table III: DNA binding activity of protein A-IN mutants using a 350 bp RSV DNA fragment containing LTR sequences

<i>Experiment 1.</i> Binding ability with mutations in the HHCC region.		
	cpm	Percent Control (%)
protein A-IN(wt)	306152 (0.1 pmoles)	100
protein A-IN(H9N)	346086	115
protein A-IN(H13N)	315192	103
protein A-IN(17–286)	327956	108
protein A	20183	0
No protein	34073	
<i>Experiment 2.</i> Binding ability of the first 63 amino acids (containing the HHCC region).		
protein A-IN(wt)	204436 (0.07 pmoles)	100
protein A-IN(1–63)	24318	0
protein A	23963	0

Conditions are described in Materials and Methods. The substrate was a 350 bp permuted LTR fragment [37] which contained covalently linked terminal sequences. Substrate input in all reactions was 1 pMole. The number in parenthesis equals pMoles bound in positive control.

Fig. 6. Location of IN alterations. The amino acid sequence of RSV IN is shown. Arrows indicate end points of deletions or sites of amino acid substitutions. The terminology used for deletions identifies the amino acids which are retained in the modified proteins. The substitutions (H9N and H13N) are denoted by the identity (e.g. H) and location (e.g. residue numbers 9 or 13) of the amino acid in the wild-type protein followed by the new amino acid (e.g. N).

Fig. 7. Site-specific endonuclease assay of protein A-IN mutants. Conditions were as described in Fig. 5.

to bind a natural 350 base pair DNA fragment which contains tandemly joined RSV LTR termini. IN is expected to bind to such viral DNA sequences with a modest preference compared to non-viral DNA [49]. The results (Experiment 1) show that the IN(H9N), IN(H13N), and IN(17–286) protein A fusions bound this fragment with the same efficiency as the wild-type control. In contrast (Experiment 2), protein A-IN(1–63), which contains only the first 63 amino acids of IN, but the entire HHCC region, showed no detectable binding to the fragment. Similar results were obtained with LTR oligonucleotides such as those used as substrates in Figure 5 (data not shown). Thus, the HHCC region, by itself, exhibits neither viral-specific nor sequence-independent DNA binding.

Site-specific DNA endonuclease activity. IN fusion proteins were tested for Mn^{2+} -dependent endonuclease activity using the U3 LTR oligonucleotide substrate. Protein A-IN(H9N), which showed wild-type DNA binding activity, had no detectable endonuclease activity (Fig. 7). Protein A-IN(H13N) showed a level of endonuclease activity similar to the protein A-IN(wt) control, but produced relatively more (about 30%) of the –3 product than the wild-type enzyme under these assay conditions. Protein A-IN(17–286) was a less active endonuclease and produced approximately equal amount of the –2 and –3 products (Fig. 7). These three HHCC mutations, which eliminate or modify the specificity of the endonuclease activity, suggest that this region contributes to the recognition of viral target sequences and/or the endonuclease activity. As might be expected, a deletion mutation which showed reduced DNA binding [*e.g.* protein A-IN(1–206)], was also negative for endonuclease activity (data are not shown).

DNA binding and endonuclease assays with non-fusion IN mutants. To test the effects of the H9N and H13N substitutions in the context of mature IN protein, these mutations were transferred to an IN expression clone (pRC23p32; [37]), and the non-fusion mutant proteins were purified and analyzed. As observed with the fusion proteins, DNA binding activities of both H9N and H13N were similar to wild-type IN (data are not shown). Figure 8A (left) shows the Mn^{2+} -dependent endonuclease activity of the non-fusion IN mutant proteins. H9N which was inactive as a fusion protein, now shows activity approximately equal to that of H13N. As can be seen on the gel, and in the quantitation shown in Figure 8B, both mutant proteins

Fig. 8. Oligonucleotide cleavage assay for non-fusion IN and H9N, H13N mutants. A. Conditions were as described in Fig. 5. Reactions with Mn^{2+} were for 15 min, and with Mg^{2+} for 30 min. B. Quantitation of cleavage products from a 30 min reaction in the presence of Mn^{2+} . Each lane in the gel was sliced into 3 mm sections and ^{32}P cpm determined from Cerenkov radiation in isolated gel slices. Numbers in bars indicate percent of substrate (S) or cleaved (–2 and –3) products of total counts per gel lane.

were approximately half as efficient as the wild-type protein, but the proportions of –2 and –3 products were similar. All three enzymes also produced a small percentage of products that migrate as a faint ladder above the substrate band. Nucleotide sequence determinations (not shown) indicate that these are products of recombination in which the 3' ends of the oligonucleotide substrates which were cleaved after the CA sequence (Fig. 5B), are joined to a cleaved internal site in a second oligonucleotide molecule. This cutting and joining reaction mediated by IN alone is consistent with our independent analyses which show that ASLV IN can catalyze integrative recombination [18].

As has been shown previously, in the presence of Mg^{2+} (Fig. 8A, right) the wild-type IN protein is less active but much more specific for cleavage at the –2 site. However, in these assays the mutant proteins were barely active, producing less than *ca.* 5% of the –2 product formed by the wild-type enzyme. Thus, the reaction which most closely resembles the *in vivo* LTR cleavage, seems to be most sensitive to alterations in the HHCC region.

DISCUSSION

Deduced amino acid sequences reveal two conserved regions in the IN proteins of retroviruses and retrotransposons; one, a region with four conserved residues HHCC, and a second, with a

conserved D(35)E motif. Except for the defining HHCC residues and a conserved hydrophobic amino acid, the residues in the HHCC regions are poorly conserved, while those in the D(35)E region are strongly conserved. The ASLV IN virion protein has been shown to possess several independently distinguishable functions *in vitro*: These include DNA binding, cutting, and joining [13,18]. The goal of this study was to construct altered proteins in order to locate sequences involved in DNA binding, and to investigate the role of the conserved HHCC motif in the DNA binding and endonuclease activities of ASLV IN.

The HHCC region. The possibility that this region may bind a Zn^{++} ion has been noted previously [28]. However, the HHCC region of RSV differs from the prototype DNA binding Zn finger (Xfin-31) of *Xenopus laevis* [51] and the nucleocapsid protein of retroviruses (50) in several respects; the length of the HHCC region in the IN proteins (measured between H13 and C37; Fig. 2) is usually either 23 or 32 residues, about twice the size of Xfin-31 and very much longer than that of the nucleocapsid protein; the order of the conserved residues HHCC is different from that found in any of the characterized Zn fingers. Since the formation of a Zn finger depends on the three-dimensional arrangement of the four coordinating residues, rather than on their order in the primary sequence, it seemed reasonable to expect that the HHCC sequence could form a Zn finger similar to that formed by the CCHH sequence.

Along with the HHCC residues, a hydrophobic group (residue 19, L or similar) is strongly conserved in this region of IN. Three other hydrophobic groups are more weakly conserved; L2, residue 23–25 (aromatic), and V33–34. Hydrophobic regions form the cores of globular proteins [52] and tend to be important in defining the shape of proteins [53]. Thus, we could speculate that the conserved location of the hydrophobic residues in the HHCC region is associated with a conserved folding pattern. However, the position of the HH residues varies within the presumed folding pattern and sequence conservation is poor between positions 19 and 33. The structure predictions, which are based on the circular dichroism (CD) data for RSV IN [36], showed no consistent secondary structure for the HHCC region in different IN proteins. In some cases, these predictions were sensitive to small changes in the CD data and may not be accurate; nevertheless the structural homology between the different HHCC regions appears small. In any case, the structures predicted for the IN HHCC regions are different from that of the Xfin-31 Zn finger [54].

The structure in the HHCC region could be determined by factors other than the sequence alone; factors such as Zn binding. For peptides derived from a retroviral NC protein, Zn is required to produce the defined secondary structure [50]. However for other Zn fingers, secondary structure predictions [55,51] confirmed by subsequent structure determination [54] or modelling [56] show that the peptide legs will fold to close to the final structure without Zn.

Our starting hypothesis was that the N-terminal HHCC region may form a 'finger' structure and, like such structures in the well-characterized transcription regulatory proteins, this region might be important in DNA binding. The results showed that the deletion of 34 [protein A-IN(1–252)] or 80 [protein A-IN(1–206)] amino acids from the C-terminus reduced this binding by a similar (ca. 60%) amount. Thus we conclude that the C-terminus must contribute in some way to binding, possibly through indirect conformational affects. Larger deletions from the C-terminus, which nevertheless should leave the HHCC

region intact [protein A-IN(1–149), (1–105), (1–63)], showed no detectable DNA binding.

We also made more modest changes designed specifically to disrupt a Zn finger [IN(17–286), (H9N), (H13N)]. Somewhat surprisingly, we found that these altered proteins bound DNA as well or better than the wild-type IN, whether the substrate was a d(AT) copolymer or a 350 base pair fragment of DNA that contained RSV LTR sequences. When the latter substrate was incubated with a truncated protein which retained an intact HHCC region, [protein A-IN(1–63)], neither LTR-specific nor sequence-independent binding was observed. Although these binding assays are limited in sensitivity, the results are clearly inconsistent with the simplest hypothesis, that the N-terminal HHCC region is a DNA binding domain similar to that described for transcription factors.

When these same HHCC mutant IN fusion proteins were tested for endonuclease activity we found that the H13N substitution was less active than wild-type, and the N-terminal deletion [IN(17–286)] was both less active and less specific. However, even with the N-terminal deletion which lacks the two histidine residues of the HHCC region, endonuclease activity was still detectable. As non-fusion proteins, the H9N and H13N substitutions exhibited properties similar to one another. Both were approximately half as active as wild-type IN with Mn^{2+} , but the pattern of cutting of LTR sequences seemed not to be affected. Both of the mutant enzymes also exhibited a Mn^{2+} -dependent joining reaction, again, with about half the efficiency of wild-type. In the presence of Mg^{2+} , where the wild-type IN is less active but apparently more specific for the correct (–2) LTR cleavage site, the H9N and H13N mutants were virtually inactive. The fact that the substitution mutants showed similar properties in Mg^{2+} and Mn^{2+} supports the notion that the H9 and H13 residues contribute to the same structural element and that this element may be involved in the correct recognition of LTR sequences in the metal-dependent integration reaction. The strong DNA binding function of IN must, however, be encoded in another domain(s). We suggest below that the D(35)E region is a likely candidate.

The D(35)E region. Sequence alignments have revealed a highly conserved 50 residue section of IN, the D(35)E region. Structure prediction programs suggest a conserved folding pattern. Conserved residues, both positively charged and hydrophobic (57), could be involved in DNA interactions. Since metal ions (Mg^{2+}/Mn^{2+}) are required for the DNA cutting, and joining activities of IN proteins, they might also be bound by conserved residues. The two residues which could serve this role are the conserved D121 and E157 which define the D(35)E region. The conserved D121 is contained in the sequence TDNG or similar (with variability allowed in the 3rd position). A D is found in a similar environment in all eukaryotic protein kinases [58]. Furthermore, in the catalytic subunit of cAMP dependent protein kinase, residues D184 (TDFG context), as well as E91, are protected by MgATP against cross linking. This suggests that D184 and E91 are close to the active site [58], and that these amino acids play a role in phosphate transfer.

The potential for DNA binding, together with MgATP and phosphate transfer associated with D and E in other proteins, suggests that the D(35)E sequence in IN takes part in the binding, cutting and transfer of DNA strands. The results of our DNA binding studies with ASLV IN are consistent with such a function; the C-terminal deletions which entered this region [IN(1–149)], or removed it entirely [IN(1–105), IN(1–63)], showed no

significant DNA binding. However other, more direct, analyses will be needed to test this hypothesis and site-directed mutagenesis studies are currently underway.

The similarity in the immediate vicinity of E157, between IN proteins and other DNA cutting and joining enzymes, has been noted before [60]. However, the whole D(35)E region of IN, and not just the residues near E157, is conserved between IN proteins and a number of bacterial IS elements (Fig. 3; [27]). This conservation has also been identified in an independent computer study by Fayet *et al.* [61]. Sequence similarities between the ends of ASLV LTRs and IS elements were noted over a decade ago [12] and a recent comparison of IS inverted terminal repeats [47] shows a high frequency (9 of 26) at the tips of these elements of the sequence 5'TG...CA3'; the same sequence is found at the proviral DNA ends of retroviruses and is part of the *cis*-acting region required for integration. Mechanisms involved in cutting and joining to host/target DNA may also be similar since a majority of the known IS elements, like retroviruses, generate small direct repeats of target DNA during insertion. The approximate definition of functional boundaries and identification of conserved amino acids will serve as a useful guide for further mutational studies and structure/function analyses of retroviral IN and other prokaryotic and eukaryotic recombination enzymes.

ACKNOWLEDGMENTS

One of us (JPGM) would like to thank Michael Gribskov, Crystallography Laboratory, NCI-FCRDC, Frederick, Maryland, and Joel Sussman, Department of Structural Chemistry, Weizmann Institute of Science, Rehovot, Israel, for helpful discussions. We acknowledge the National Cancer Institute for allocation of computing time and staff support at the Advanced Scientific Computing Laboratory of the Frederick Cancer Research and Development Center. Doug Markham and Jenny Glusker of the Fox Chase Cancer Center and Jonathan Leis of Case Western Reserve University, provided many helpful suggestions for improvements in the manuscript. We would also like to acknowledge the contribution of R. Terry who constructed the *Bal31* IN deletions used to produce our protein A fusions.

This work was supported by United States Public Health Service Grants CA-49042, CA-48703, RR-05539, CA-06927, a grant from the Pew Charitable Trust and an appropriation from the Commonwealth of Pennsylvania. Support for R.A.K. and J.K. was provided by the W.W. Smith Charitable Trust.

REFERENCES

- Varmus, H. E. and Swanstrom, R. (1982) In R. Weiss, N. Teich, H. Varmus and J. Coffin (eds.), *RNA Tumor Viruses*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- Donehower, L. A. and Varmus, H. E. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 6461–6465.
- Hippenmeyer, P. J. and Grandgenett, D. P. (1984) *Virology*, **137**, 358–370.
- Schwartzberg, P. *et al.* (1984a) *Cell*, **37**, 1043–1052.
- Brown, P. O. *et al.* (1987) *Cell*, **49**, 347–356.
- Fujiwara, T. and Mizuuchi, K. (1988) *Cell*, **54**, 497–504.
- Colicelli, J. and Goff, S. P. (1985) *Cell*, **42**, 573–580.
- Colicelli, J. and Goff, S. P. (1988) *J. Mol. Biol.*, **199**, 47–59.
- Cobrinik, D. *et al.* (1987) *J. Virol.*, **61**, 1999–2008.
- Shih, C.-C. *et al.* (1988) *Cell*, **53**, 531–537.
- Temin, H. (1980) *Cell*, **21**, 599–600.
- Ju, G. and Skalka, A. M. (1980) *Cell*, **223**, 379–386.
- Skalka, A. M. (1988) In R. Kucherlapati and G. R. Smith (eds.), *Genetic Recombination*. ASM Press.
- Varmus, H. E. and Brown, P. (1989) In D. E. Berg and M. M. Howe (eds.), *Mobile DNA*. American Society of Microbiology, Washington, DC.
- Grandgenett, D. P. *et al.* (1985) *J. Biol. Chem.*, **260**, 8243–8249.
- Alexander, F. *et al.* (1987) *J. Virol.*, **61**, 534–542.
- Katz, R. A. and Skalka, A. M. (1988) *Virology*, **62**, 528–533.
- Katz, R. A. *et al.* (1990) *Cell*, **63**, 87–95.
- Katzman, M. *et al.* (1989) *J. Virol.*, **63**, 5319–5327.
- Roth, M. J. *et al.* (1989) *Cell*, **58**, 47–54.
- Craigie, R. *et al.* (1990) *Cell*, **62**, 829–837.
- Duyk, G. *et al.* (1985) *J. Virol.*, **56**, 589–599.
- Grandgenett, D. P. and Vora, A. C. (1985) *Nucl. Acids Res.*, **13**, 6205–6221.
- Katzman, M. *et al.* (1990) Submitted for publication.
- Wang, J. C. (1985) *Annu. Rev. Biochem.*, **54**, 665–697.
- Misra, R. K. *et al.* (1982) *J. Virol.*, **44**, 330–343.
- Mack, J. P. G. (1990) In M. Cassman (ed.), *Fourth Meeting: Structures of AIDS-related systems and their application to targeted drug design*. National Institute of Health, Bethesda, Maryland.
- Johnson, M. S. *et al.* (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7648–7652.
- Devereux, *et al.* (1984) *Nucl. Acids Res.*, **12**, 387–395.
- Myers, G. *et al.* Human retroviruses and AIDS 1989: A compilation and analysis of nucleic and amino acid sequences. Los Alamos National Laboratory, Los Alamos, New Mexico, 87545, USA.
- Gribskov, M. Profile Analysis, GCG compatible version, Release 4.2, March 8, 1989.
- Hein, J. (1990) In R. F. Doolittle (ed.), *Methods in Enzymology*. Academic Press, New York, Vol. 183, pp. 626–645.
- Gribskov, M. *et al.* (1986) *Nucl. Acids Res.*, **14**, 327–334.
- Chou, P. Y. and Fasman, G. D. (1978) *Adv. Enzymol.*, **47**, 45–147.
- Garnier, J. *et al.* (1978) *J. Mol. Biol.*, **120**, 97–120.
- Lin, T. *et al.* (1989) *Proteins*, **5**, 159–165.
- Terry, R. *et al.* (1988) *J. Virol.*, **62**, 2358–2365.
- Casadaban, M. J. and Cohen, S. N. (1980) *J. Mol. Biol.*, **138**, 179–207.
- Crowl, R. *et al.* (1985) *Gene*, **38**, 31–38.
- Laemmli, U. K. (1970) *Nature (London)*, **227**, 680–685.
- DeChiara, T. M. *et al.* (1986) *Methods in Enzymology*, Part C, **119**, 403–415.
- Katz, R. A. *et al.* (1986) *Gene*, **50**, 361–369.
- Maxam, A. M. and Gilbert, W. (1980) In S. P. Colowick and N. O. Kaplan (eds.), *Methods in Enzymology*. Academic Press, New York, Vol. 65, pp. 499–559.
- Dayhoff, M. O. *et al.* (1979) In M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Suppl. 3, p. 345.
- George, D. G. *et al.* (1990). In R. F. Doolittle (ed.), *Methods in Enzymology*, Vol. 183, pp. 333–351.
- Zagury, J. F. *et al.* (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 5941–5945.
- Galas, D. J. and Chandler, M. (1989) In D. E. Berg and M. M. Howe (eds.), *Mobile DNA*. American Society for Microbiology, Academic Press, Washington, DC, pp. 109–162.
- Duyk, G. *et al.* (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 6745–6749.
- Knaus, R. J. *et al.* (1984) *Biochemistry*, **23**, 350–359.
- Summers, M. F. *et al.* (1990) *Biochemistry*, **29**, 329–340.
- Gibson, T. J. *et al.* (1988) *Protein Eng.*, **2**, 209–218.
- Kauzmann, W. (1959) In C. B. Anfinsen, M. L., Anson, K. Bailey and J. T. Edsall (eds.), *Advances in Protein Chemistry*. Academic Press, New York, Vol. 14, pp. 1–63.
- Bowie, J. U. *et al.* (1990) *Science*, **247**, 1306–1310.
- Lee Min, S. *et al.* (1989) *Science*, **245**, 635–637.
- Berg, J. M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 99–102.
- Green, L. M. and Berg, J. M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 4047–4051.
- Merrill, B. M. *et al.* (1988) *J. Biol. Chem.*, **263**, 3307–3313.
- Buechler, J. A. and Taylor, S. S. (1988) *Biochemistry*, **27**, 7356–7361.
- Buechler, J. A. and Taylor, S. S. (1989) *Biochemistry*, **28**, 2065–2070.
- Doolittle, R. F. *et al.* (1989) *Q. Rev. Biol.*, **64**, 1–29.
- Fayet, O. *et al.* (1990) *Mol. Microbiol.*, **4**, in press.