Burdock, a novel retrotransposon in *Drosophila melanogaster*, integrates into the coding region of the *cut* locus

N.A. Ponomarenko[a,b,*], V.M. Bannikov[a], V.A. Anashchenko[a], N.A. Tchurikov[a]

[a]*Engelhardt Institute of Molecular Biology, Russian Academy of Science, Vavilov str. 32, Moscow 117334, Russia*
[b]*Centre for Medical Studies, University of Oslo, Vavilov str. 34/5, Moscow 117334, Russia*

Received 19 June 1997

Abstract **The *burdock* element is known to be the 2.6-kb insertion into the same region of the *cut* locus in 12 independently obtained *ct*-lethal mutants. Here we have determined the complete sequences of this insertion and of the hot spot region. It was found that the *burdock* is a short retrotransposon with long terminal repeats and a single open reading frame (ORF). The polypeptide encoded by the *burdock* ORF contains two adjacent regions homologous to the *gag* and *pol* polyproteins of the *gypsy* mobile element. The *burdock* insertion interrupts the short ORF of the *cut* locus. The target site sequence of the *burdock* insertions is similar to the *Drosophila* topoisomerase II cleavage site.**

Key words: Mobile element; *burdock*; *cut* locus; Topoisomerase II; *Drosophila melanogaster*

1. Introduction

Mobile genetic elements are an important component of eukaryotic genomes, encompassing several percent of the *Drosophila* total DNA and account for up to 90% of the spontaneous mutations revealed by classical genetic methods. They can be classified according to their structure and presumed mechanism of transposition, and fall into two main classes: elements that transpose by reverse transcription of an RNA intermediate (retroposons), and elements that transpose directly from DNA to DNA. One of two types of retroposons are elements that resemble retroviruses in having long terminal direct repeats (LTR) and open reading frames (ORF) with the potential to code for polypeptides similar to viral *gag*, *pol* and, in some cases, *env* proteins. In addition to the conservation in the organization of the coding region, these elements also contain nucleotide sequences homologous to the tRNA primer binding site and purine-rich sequences, both necessary for the initiation of DNA synthesis. A novel mobile element of *Drosophila* (*burdock*) was detected in the course of the molecular analysis of derivatives of the unstable strain ct^{MR2} [1]. In 12 *ct*-lethals that arose independently, the genomic Southern blot analysis revealed the insertion of a fragment of about 2.6 kb in length into the −116 region of the *cut* locus (Fig. 1b). This fragment from the mutant strain ct^{MR21K1} was cloned. Using in situ hybridisation and physical mapping this insertion was shown to be a novel mobile element [2].

In the present study we have determined the nucleotide sequences of this mobile element and the region of its insertion in the *cut* locus of *Drosophila melanogaster*.

*Corresponding author. Fax: (7) 095-135-1405

2. Materials and methods

2.1. Molecular cloning and restriction analysis

*Eco*RI DNA fragments 1.57 and 4.0 kb in length were subcloned into pUC12. Some subfragments were recloned into M13 phages. Cloning and restriction analysis were performed as described earlier [1]. Plasmid DNA was amplified and isolated as described [3].

2.2. Isolation of DNA from phage M13

Amplification and isolation of single-stranded DNA from phage M13 was carried out as described [4] with following modifications: phage from a plaque was inoculated by pricking into 2 ml of exponential phase JM103 cell culture and grown for 3 h at 37°C; 0.1 ml of the obtained culture and 1 ml of JM103 cell culture were added to 9 ml of warm 2×YT medium and grown for 6 h at 37°C; the culture was centrifuged 3 times at 4000 rpm for 10 min; a solution of 3 M NaCl, 25% PEG was added to the obtained supernatant (1:7 vol.), incubated for 15 min at room temperature, and centrifuged; the precipitate was dissolved in 1 ml of 0.1 M NaCl and centrifuged for 2 min; the supernatant was removed and the procedure of phage precipitation with the solution of 3 M NaCl, 25% PEG was repeated; the obtained precipitate was dissolved in 100 μl of TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) and, after 100 μl of NEST was added, the solution was purified with a 1:1 phenol/chloroform mixture, pH 8.0; DNA was precipitated with 2 vol. of ethanol in the presence of 1:10 vol. of 3 M sodium acetate, pH 5.2; the obtained precipitate was washed 3 times with 70% ethanol containing 0.1 M NaCl and dissolved in 0.1 TE. To obtain the replicative form of the M13 phage DNA, we used the method of plasmid DNA isolation [3].

2.3. DNA sequencing

DNA primary structure was determined on both strands by Sanger's method [5].

3. Results

A λ clone containing a copy of the mobile element and adjacent genome sequences was isolated by screening the genomic library from the strain ct^{MR21K1}. The *burdock* element is known to integrate into a 1.57 kb *Eco*RI fragment and to lack internal *Eco*RI sites (Fig. 1b). Therefore, an approximately 4 kb *Eco*RI fragment from the λ clone containing *burdock* and adjacent regions of the locus was recloned into pUC12 (p4.0). By a similar procedure a 1.57 kb *Eco*RI fragment lacking the mobile element (p1.57) was isolated from DNA of the wild-type strain (Oregon RC). Fig. 1c demonstrates the physical maps of clones and *burdock* localisation obtained by restriction analysis and cross-hybridisation.

The subfragments of plasmids p4.0 and p1.57 were recloned into phages M13mp10 and M13mp11 for sequencing.

In Fig. 2 the nucleotide and putative amino acid sequences of the mobile element are presented. It appeared that *burdock* is an unusually small retrotransposon 2569 bp in length. It possesses 275-bp-long terminal repeats (LTR) lacking sequence homology with any known LTRs. The *burdock* LTRs are flanked with short AgTTA-TAAtT inverted repeats

PII S0014-5793(97)00844-2

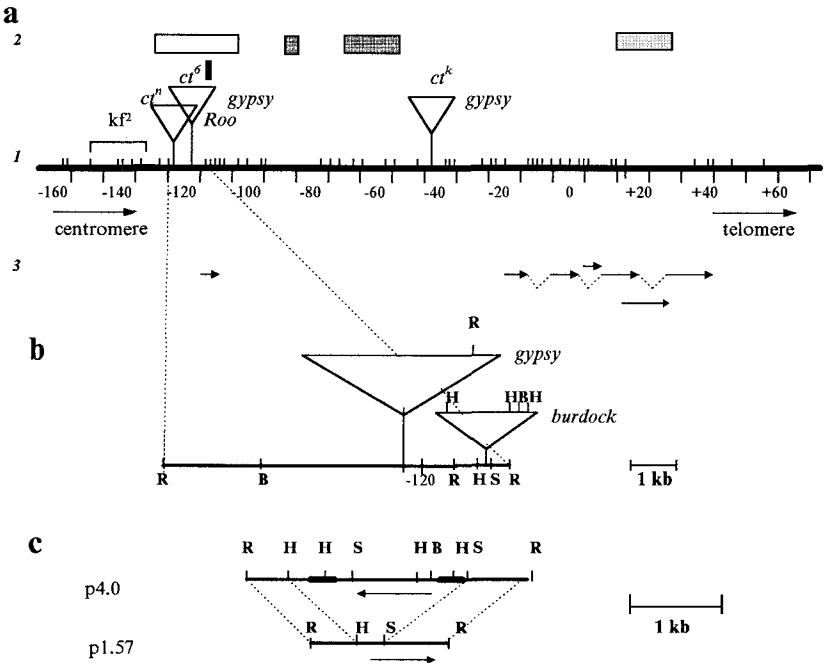

Fig. 1. a: Physical map of the *cut* locus [4]. 1: *Eco*RI map of the cloned DNA segment from the Oregon R strain; the nature and localization of the classic *ct*-mutants are indicated; the scale is in kb. 2: Locations of the known *ct*-mutations: viable (open bars) and lethal (gray bars) mutations caused by the *gypsy* insertions; hatched bars indicates *ct*-lethals not associated with mobile elements [16]; solid bar indicates the site of *burdock* insertions [1] 3: Arrows indicate localization and direction of known transcripts in the locus [5,18,19]. b: Physical map of the *burdock* insertion region of the strain ct^{MR1K1} [1]. c: Restriction maps of p4.0 containing *burdock* and its insertion region and p1.57 containing only the same region of the *cut* locus. The *burdock* LTRs are indicated by solid bars; arrows correspond to ORFs; the direction of transcription is shown in the locus and in the mobile element; R, *Eco*RI; H, *Hin*dIII; S, *Sal*I; B, *Bam*HI.

and carry two putative polyadenylation signals (at position 226 and 241, Fig. 2). No sequence identical to any previously described for mobile elements RNA start site was found in the nucleotide sequence of the *burdock* LTR. There are the $tRNA^{Lys}$ binding site downstream from 5′-LTR sequence and the polypurine tract upstream of 3′-LTR sequence.

Burdock has one ORF that could specify a protein of 238 amino acids. A computer-assisted search among SWISSPROT database (release 34) using the GENEBEE program [6] revealed the homology between the putative *burdock* polypeptide and the proteins encoded by *gypsy* retrotransposon [7]. Interestingly, 120-amino-acid residues 53–173 are homologous to the corresponding part of *gypsy* ORF1, whereas the neighbouring 65 amino acids are homologous to the carboxyl end of ORF2 of this mobile element (Fig. 3).

The insertion of the *burdock* element to the *cut* locus leads to duplication of the tetranucleotide TATA and interrupts the ORF encoding a polypeptide of 220 amino acids in length. Fig. 4 presents the nucleotide and putative amino acid sequences of the corresponding region of the *cut* locus. Data of the genomic blot analysis revealed that in all 12 *ct*-lethals resultant from the *burdock* insertion transcription within mobile element is in the opposite direction of that within the *cut* locus.

It was shown that the sequence surrounding the 'hot spot' of the *burdock* insertion is similar to the *Drosophila* topoisomerase II cleavage site. This region contains two overlapping consensus sequences recognised by topoisomerase II: pentadecamer GTN(T/A)A(C/T)ATTNATNN(A/G) and the alternating purine–pyrimidine sequence [8–10]. The integration site of the mobile element itself is incorporated into both consensus sequences (Fig. 5).

4. Discussion

Molecular analysis of the given copy of the *burdock* retrotransposon revealed a series of structural peculiarities in this novel mobile element.

Some structural features, such as the inverted AgTTATAAtT repeats flanking the LTRs, the $tRNA^{Lys}$ binding site, and the high degree of homology between the protein products (up to 76%) suggest a relationship between *burdock* and *gypsy* elements and possibly indicate their common origin. However, there is no sequence homology somewhere else except coding region and *burdock* contains an approximately 1 kb non-coding sequence instead of ORF3 in *gypsy* that encodes the proteins of the viral envelope [11].

No promoter region identical to any of those previously described for mobile elements [12–15] was found in *burdock* LTR sequence. In addition, the absence of ACA(G/C) or CGTG tetranucleotide which is known as a downstream +30 element of retrotransposon promoters, suggest that the *burdock* possesses RNA start-site distinct from anyone described, earlier. Therefore, we propose to study *burdock* transcription and its promoter region.

The mechanism of transposition of this copy of the element remains unclear because of the absence of domains coding for RNAse H, reverse transcriptase, protease, and integrase in the *burdock* coding region. The character of homology between *burdock* and *gypsy* suggests that this *burdock* sequence resulted from the deletion of the approximately 3.7 kb fragment in the 1085–1095 region which, like the corresponding *gypsy* region, codes for the domains required for transposition. Therefore, it seems probable that a complete *burdock* copy, containing all sequences needed for transposition and capable

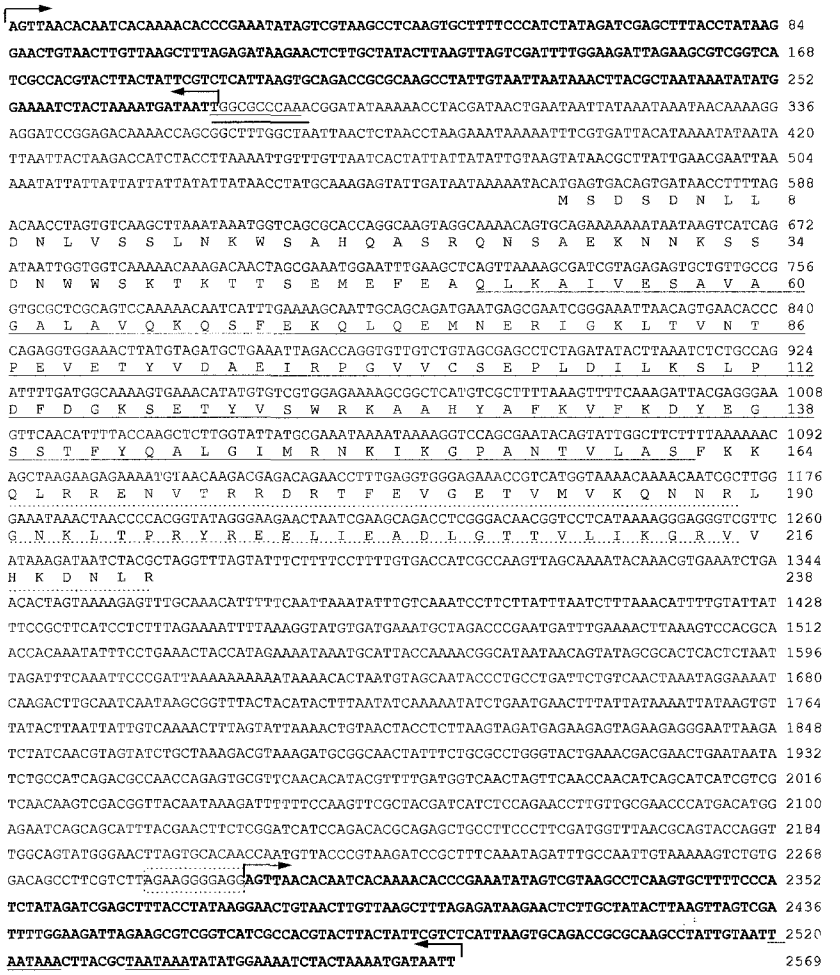

Fig. 2. Nucleotide and amino acid sequences of *burdock* element. LTRs are shown in boldface. The arrows indicate the starting and ending nucleotides of the LTRs; the tRNA binding site is underlined twice; the polypurine tract is included in the dashed frame; polyadenylation signals are underlined. The amino acid region homologous to *gypsy* ORF1 is underlined and dotted line indicates that homologous to *gypsy* ORF2.

of ensuring integration of the deleted variant, is present in the genome. Studying the other *burdock* copies of the *Drosophila* genome will allow us to solve this issue.

The *burdock* insertion interrupts the ORF that codes for a putative polypeptide of 220-amino-acid residues in the transcribed region of the *cut* locus. Since this sequence does not contain the methionine codon at the origin, it is likely to be a exon within the longer transcription unit. It is of interest that mutations $ct^{81/1}$ and ct^{L-18} [16] resulted from the *gypsy* insertion into the same 0.35 kb *Hin*dIII–*Sal*I fragment of the cut locus are viable. The direction of transcription in mobile element is the same as in the locus. Thus, the lethal phenotype that arises due to the insertion of *burdock* apparently emerged as a result of insertion in a defined orientation.

It was found that the sequence around the hot spot of the *burdock* insertion corresponds to the topoisomerase II cleav-

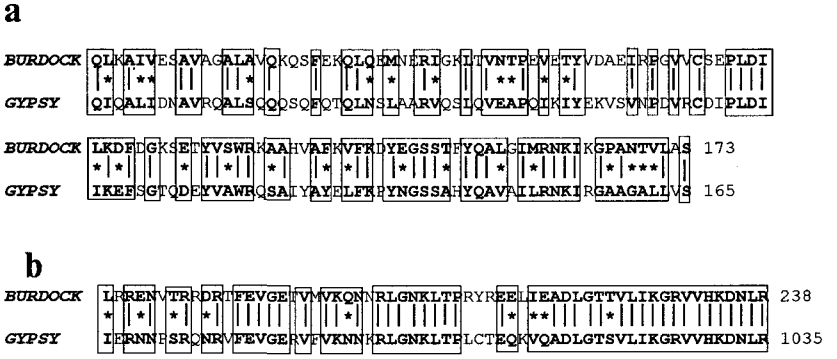

Fig. 3. Alignment of the amino acid sequences encoded by burdock ORF and gypsy ORF1 (a) and ORF2 (b). Identical (|) and similar (*) residues are shown in boldface and framed according to [20].

```
GAATTCTTACCCCGCACACTCCAGTCTCCATACAGAGATGGATCATGATTGCAACAATGATGACGTTGAACTGCTCTCCGATTC 84
GGAATTAGAATCAGAATCAGAATCAGAATCACTACAGTGCCAAAGGAATGACCACCGCTCAACCGATTGCCGTACCTCTCCGGG 168
TGTCATATAATAATCGTCTGTATTTCCTTGATCAATTGTATGTCGTCGCTTCTTGCTAGATTGCTTCTTTCTTAACACCTTACT 252
GCGACCCATTTTCCGTATTTCCGTATGTGTCTTTTAACTGTTGATAGTATATTAATATCTGAGGTTGTAATTGTACTGTCGGAA 336
ATTTTAAAATGGAGTGAAGTCTGCAACATCAAACTGACAGGTTTCTCAAAACAAAAAATCTGTACCATAATTTGTATCTTACTG 420
TACATTACATGTACCTTATGTTCCATAAGCGTTTTCGTCAAAAATAGGGCCTACCGAATACCCCGAAATCGATTAGTGCGGGCC 504
AACAAGTCGATAAGATCATAGAGGAGTTATAGGTCTTGCATATTAGCTGGAAGCTCTTCTCCGAAGAGATCGATTGTGTGTGGC 588
TTCCATTGAAAATTGGCTACCGACCACACCATATGGGGCAAAACTAGATCATAGTATCCATTTAATGCGTGTAGATCGAAGGGC 672
TAAGATAGCCGACAGTTAGTTGGAGAACCTTTCAGTTAGTGGGTTAAGCTTAAAGGCGACAATCTCGTGCACCCGAAACCCGCT 756
ACAAAAACGCGGAAAAAACGTATTAAATACAAATTTCCTGGTAATAAAACAGAGAGACAACTCTGGCCACTGGCTGCCAAAATA 840
                                                D  R  E  T  T  L  A  T  G  C  Q  N   12
AAGAAATAAATCTTAACGCCCAGAGGGTCAGAGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAAAGACAGCGAATCGGAATCG 924
K  E  I  N  L  N  A  Q  R  V  R  E  R  E  R  E  R  E  R  E  R  E  R  Q  R  I  G  I   38
GAGTGACTAACGTATCGCCGCTGCTTCCAGCCTTTATTGTTACTTTTGTGTATATATTCTTGTATCAGCGTGTAAATCTCGCTC 1008
G  V  T  N  V  S  P  L  L  P  A  F  I  V  T  F  V  Y  I  F  L  Y  Q  R  V  N  L  A   64
ACACACCAGATGAAATACGGAGAGTCGGCCGAAAACCGAACTCAAGTCGACAGAATTTTTATTGGAAATCTGTACGCAGAAATA 1092
H  T  P  D  E  I  R  R  V  G  R  K  P  N  S  S  R  Q  N  F  Y  W  K  S  V  R  R  N   90
GTCATCATCCCATTCCAACCCCGTCCATCCCCATCGTCGTCATAATGATCTTTGGATTCGGATTCGTATTCGGATTCGGATTCA 1176
S  H  H  P  I  P  T  P  S  I  P  I  V  V  I  M  I  F  G  F  G  F  V  F  G  F  G  F   116
GCCGCATCTTCATCTGCCGCCTGATCTTTTTCTTGTTTCTGATGCCTCAGATGCATATCGCAGTTCGCGCTGGCAATAAAAATA 1260
S  R  I  F  I  C  R  L  I  F  F  L  F  L  M  P  Q  M  H  I  A  V  R  A  G  N  L  N   142
TATATGAGCTGTTGATGACTGCATATAAAAACAGGATGGAATGCTATTGTTGCTCGACTATCGAGTACCCGTTAGTCAGAGAAA 1344
I  Y  E  L  L  M  T  A  Y  K  N  R  M  E  C  Y  C  C  S  T  I  E  Y  P  L  V  R  E   168
ATTTTTTTAAAAGTGTGGACGGGAGAGGCATGCTAAGTATCGAATCTGATCAAGAATATAATTATTTAATATGGTCAGAAAGGC 1428
N  F  F  K  S  V  D  G  R  G  M  L  S  I  E  S  D  Q  E  Y  N  Y  L  I  W  S  R  R   194
TTAATTCCACATCTTACTTTGTTTTTCAGCCAATCTAGTATACCCTTTTACTTCACGTGCAACGGGTATAATTATACGAAACTG 1512
L  N  S  T  S  Y  F  V  F  Q  P  I                                                    220
CCAGTGAACAAACCCTCGGATATACTGTGATATATGCGTGCCGATGCAGTTGACTCCGGGACCAAAGTCAAGGAATTC       1590
```

Fig. 4. Nucleotide and amino acid sequences of *burdock* insertion region of the *cut* locus. The duplicated sequence is shown in boldface.

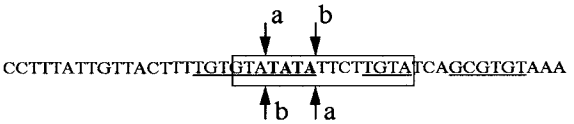

Fig. 5. Nucleotide sequence of the cut locus corresponding to the topoisomerase II cleavage site. TOPOII consensus is framed; alternating purine–pyrimidine sequence is underlined. The duplicated sequence is shown in boldface; arrows indicate two equally possible variants (a and b) of DNA breakage during *burdock* integration.

age site. Earlier, similar sequences were found near the site of integration of human immunodeficiency virus (HIV) [17]. If topoisomerase II could modify a chromosome conformation, we suppose that it may be involved in the integration of retroelements into some genomic regions.

Acknowledgements: We thank Dr. V. Zakharyev and Dr. E. Nahon for advice and Dr. A. Pellisson for discussion. This work was supported by INTAS (Grant 94-06345), Russian Foundation for Basic Research (grant 97-04-49755) and the Russian State program 'Frontiers in Genetics' (Grant 003a).

References

[1] Tchurikov, N.A., Gerasimova, T.G., Johnson, T.K., Barbakar, N.I., Kenzior, A.L. and Georgiev, G.P. (1989) Mol Gen Genet 219, 241–248.
[2] Gerasimova, T.G., Obolenkova, L.A., Kiselev, S.L. and Tchurikov, N.A. (1990) Dokl Akad Nauk SSSR 311, 474–476.
[3] Lee, S. and Rasheed, S. (1990) Biotechniques 9, 676–679.
[4] Messing, J. (1983) Methods Enzymol 101, 20–78.
[5] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc Natl Acad Sci USA 74, 5463–5467.
[6] Brodski, L.I. and Vassiliev, A.V. (1992) Amer Math Soc DIMACS 8, 127–140.
[7] Marlor, R.L., Parkhurst, S.M. and Corces, V.G. (1986) Mol Cell Biol 6, 1129–1134.
[8] Sander, M. and Hsieh, T. (1985) Nucl Acids Res 13, 1057–1072.
[9] Lee, M.P., Sander, M. and Hsieh, T. (1982) J Biol Chem 264, 21779–21787.
[10] Spitzner, J.R., Chung, I.K. and Muller, M.T. (1990) Nucl Acids Res 18, 1–11.
[11] Kim, A., Terzian, Ch., Santamaria, P., Pellisson, A., Prud'homme, N. and Bucheton, A. (1994) Proc Natl Acad Sci USA 91, 1285–1289.
[12] Flavell, A.J., Lewis, R., Simon, M.A. and Rubin, M.G. (1981) Nucl Acids Res 9, 6279–6291.
[13] Jarell, K.A. and Meselson, M. (1991) Proc Natl Acad Sci USA 81, 102–104.
[14] Micard, D., Couderc, J.L., Sobrier, M.L., Giraud, G. and Dastique, B. (1988) Nucl Acids Res 16, 455–470.
[15] Inouye, S., Hattori, K., Yuki, S. and Saigo, K. (1986) Nucl Acids Res 14, 4765–4778.
[16] Jack, J.M. (1985) Cell 42, 869–876.
[17] Howard, M.T. and Griffith, J.D. (1993) J Mol Biol 232, 1060–1068.
[18] Tchurikov, N.A., Ponomarenko, N.A. and Georgiev, G.P. (1988) Dokl Akad Nauk SSSR 303, 984–986.
[19] Blochlinger, K., Bodmer, R., Jack, J., Jan, J.Y. and Jan, Y.N. (1988) Nature 333, 629–634.
[20] R.M. Schwartz, and M.O. Dayhoff, in: M.O. Dayhoff (Ed.), Atlas of Protein Sequence and Structure, Vol 5, National Biomedical Research Foundation, Washington, DC, 1978, pp. 353–358.