A reappraisal of non-consensus mRNA splice sites

Ian J.Jackson
MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Received June 6, 1991; Accepted June 18, 1991

INTRODUCTION

Most protein-coding genes of higher eukaryotes are interupted by introns, which are removed from the primary RNA transcript by a process termed splicing. The bases which form the borders of introns and exons comprise the sequences essential for mRNA splicing, carried out by a ribonucleoprotein complex, the spliceosome. Over the 13 years since the discovery of this type of intron, several thousand intron/exon junction sequences have been determined, which has allowed the compliation of consensus sequences for both the 5′ and 3′ splice sites (1–3). The recognition of the splice site by the spliceosome is of fundamental importance for the accurate expression of genes. The recognition of splice sites in DNA by computer searches is of importance for interpreting genomic sequences, and identifying exons in streches of otherwise undistiguished DNA.

The consensus splice site sequences have therefore been much examined. Compilations of splice sites over the years have used increasing numbers of sites to derive a consensus, which in general has remained unaltered. Although the splicing substrate is, of course, RNA, most analyses are carried out on DNA, so I will refer to the sequences in the DNA form (i.e. using T rather than U). The consensus sequence at the 5′ end of introns is:

(A/C)AG GTAAGT
−1 +1

in which the cleavage occurs between the G residues at positions −1 and +1. The sequence is the perfect complement of bases 4–12 from the 5′ end of the U1 snRNA component of the spliceosome:

5′ ACUUACCUG 3′

which is thought to guide the selection of the 5′ cleavage site. (The complementarity between the first C of the consensus and the G at base 12 of U1 RNA is possibly not important for splice site selection (4)).

After cleavage, the 5′ end of the intron forms a lariat structure using a 2′ phosphodiester bond to an A residue near to the 3′ end of the intron. The 3′ end of the exon is joined to the next exon, defined by a 3′ splice site consensus of:

$(Py)_n$NPyAG G
−1 +1

where cleavage is again between the two G residues.

Most positions of the both consensus sequences are not invariant; at some positions only 50–60% of splice sites actually have the consensus base (see below for a discussion of the ways in which the sites can diverge from the prototype). However, the first two and the last two bases of the intron are invariant; which gave rise to the so-called GT-AG rule for splice site selection (5).

Among the thousands of splice sites analysed by Shapiro *et al* a handful were identified which did not conform to the GT-AG rule (3). Prompted by the identification of a non-consensus splice in the murine TRP-1 gene (52) I have reassesed these previously published non-GT-AG sites, eliminated a number as clear errors or misinterpretation and identified several more. Almost all these non-consensus splices now fall into particular patterns which allow further rules for splice consensus sequences to be deduced.

A number of immunoglobulin genes appear to have non-consensus splice sites which do not fall into these patterns. However, there are problems in determining if particular immunoglobulin genes are active, and what is the actual genomic sequence from which the mRNA has been transcribed. I do not consider any immunoglobulin sequences in this review. Neither do I consider any yeast (*S. cerevisiae*) sequences. Yeast introns are spliced by a mechanism basically the same as in higher eukaryotes; but with a stronger conservation of branchpoint sequence and slightly different 5′ and 3′ consensuses (see 6 for a recent review).

ELIMINATING ERRORS OF IDENTIFICATION

I have reassessed the non-consensus sites tabulated by Shapiro *et al*, by examining the relevant database entry, by examining the original published data, and in some cases by communication with relevant labs. The sequences which were thought to be non-consensus sites but which I have eliminated are shown in Table 1. The grounds for eliminating each sequence are as follows:

5′ Splice Sites

a) *Chironomus thumni* Balbiani ring c locus gene.
The original paper (7) identifies the splice as conforming to a good consensus match. The site as shown by Shapiro *et al* has been placed one base 3′ of the correct site.

b) *Drosophila* opsin gene.
The original paper (8) has all splices conforming to the consensus. There appears to be an error in database annotation.

c) Human MHC Class II gene.
This gene is described by the authors (9) as a pseudogene, and its transcript is most likely not processed. Although it does contain a variant consensus GC sequence (see below) as the first two

Table 1. Previously published non-consensus splice sites eliminated in this review.

Genbank locus	gene	sequence	reason for elimination
5′ splice sites			
Chibrcb	Chironomus Balbiani ring c locus	GAG TAAGTT	misalligned 1 base 3′
Droopsb2	Drosophila opsin	GAT TGCCTA	database annotation error
Hummhsxa	Human MHC class II	ATG GCACTG	non-functional pseudogene
Musgfn1	Mouse nerve growth factor	GCA TCGGTG	database annotation error
Musgfn3	Mouse nerve growth factor	TGC AGAATT	database annotation error
Mushox162	Mouse Hox-1.6	CAG GGAAGG	error in original published sequence
Droantpg4	*Drosophila* Antennapedia	CAG GAAAGT	possible polymorphism
Hamcryaa1	Hamster crystallin	AGG CAAGTT	missalligned base
3′ splice sites			
Adbcg	Adenovirus	GGGGTCGTGC A	database annotation error
Chibrcg	Chironomus Balbiani ring c-locus	CGAAAGCAAT G	database annotation error
Chkmyhd	Chicken myosin heavy chain	TCCTCTGTCA A	incorrect database entry
Humplp4	Human myelin proteolipid	GTTTGTGGGC A	database annotation error
Humplp5	Human myelin proteolipid	CCTCTTTTCA T	database annotation error
Mlap531	Murine leukaemia virus insertion	CTAGTCCCGC T	database annotation error
Muscd42	Mouse CD4 T-cell antigen	GGAGACCACC A	atabase annotation error
Musgstya2	Mouse glutathione S-transferase	AGTTGCTGCA A	database and interpretation errors
Muslyt212	Mouse Lyt-2 T-cell antigen	GACCTGGACC T	interpretation error
Rattma3	Rat α-tropomyosin	AAGAGTTGAA A	interpretation error
Sehcryaa1	Mole rat α-crystalin	CCATCAAGGA A	interpretation error

Table 2. Summary of GC consensus 5′ splice sites. The sequence on the top row is the prototype sequence which is fully complementary to the 5′ end of the U1 snRNA. The gap in the sequences separates exon from intron sequence.

GENE	SEQUENCE	REFERENCE
prototype	CAG GTAAGT	
human acetylcholine receptor	AAG GCAAGG	28
human factor XII	CCG GCGAGT	29
human factor VII	CAG GCGGGG	30
human cytochrome P-450	AAG GCAAGC	31
human prothrombin	CTG GCAAGT	32
human APRT	CAG GCGAGT	33
human keratin	GAG GCAGGC	34
human superoxide dismutase	AAG GCAAGG	35
mouse superoxide dismutase	AAG GCAAG	36
hamster α-crystallin	AAG GCAAGT	37
mouse α-crystallin	AAG GCAAGT	38
mole-rat α-crystallin	AAG GCAAGT	39
hamster APRT	CAG GCGAGT	40
mouse APRT	CAG GCGAGT	41
mouse RNA polymerase II	AAG GCATGT	42
mouse RNA polymerase II	CAG GCAAGA	42
rat haem oxygenase	CAG GCAAGC	43
mouse TRP-1	AAG GCAAGT	52
pig growth hormone	CAG GCAAGT	44
bovine aspartyl protease	GAG GCAAGT	45
chicken myosin	CAG GCAAGT	46
chicken α-globin	AAG GCAAGC	47
duck α-globin	AAG GCAAGC	48
earthworm haemoglobin-c	CAA GCAAGT	49
neurospora qa repressor	TAG GCACGT	50
soybean nodulin-24	AGG GCAAGT	51

bases of the intron, the surrounding sequence is not a good consensus match and I have neverthless excluded it from further consideration.

d) Mouse NGF.

This gene occurs twice in the list of sites, for two splices. Comparison with the original paper (10) indicates that there has been a misinterpretation in the database entry, such that intron and exon sequences have been interchanged. The correctly-assigned splices have normal consensus splice sites.

e) Mouse *Hox-1.6.*

The original publication (11) does not have a consensus 5′ site, and the authors do not comment on it. However, a more recent sequence (L. Gudas personal communication) shows a T rather than G at position 2 of the intron, resulting in a consensus splice site.

f) *Drosophila Antennapedia* gene.

Analysis of cDNAs from this gene (12) reveals two splice sites, 12 bp apart. When they are alligned on the genomic sequence, the second of these has GA rather than GT at the 5′ end of the intron. However, the cDNAs analysed were from embryos of the Oregon R strain of flies, whilst the genomic sequence was derived from a different strain, Canton S. Schneuwly *et al* (12) suggest that there is a polymorphic difference between the two strains, the result being that Oregon R embryos use an additional splice site not present in Canton S. Until more information is available I exclude this splice from consideration.

g) Syrian golden hamster α-crystallin gene.

A duplication around the intron site gives ambiguity of location of this intron. If the intron site is moved 1bp 5′, the result is a GC variant consensus splice, identical to that found in the mouse α-crystallin gene, and I include this hamster gene in Table 2.

The remaining 17 sites described in Shapiro et al, plus the hamster α-crystallin gene described above, the mole-rat α-crystallin below and a further 7 sites which I have identified through literature searches are shown in Table 2 and discussed in the next section.

3′ Sites

a) Adenovirus.

The sequence given as a 3′ splice site, is not in fact a site. It appears to be an error in database annotation; the site given is found at base 24791 of the sequence, but the actual splice site is at 24971 (13). The sequence at this site is a consensus 3′ AG.

b) *Chironomus thummi* Balbiani ring c-locus protein.

The site described is not a 3′ splice site. There appears to have been misinterpretation by the database, as the original paper (7) shows a change in protein domains, not an intron, at the site listed.

c) Chicken myosin heavy chain gene.

In this case there appears to have been incorrect entry of sequence into the database. The original paper (14) has a G, which is absent

Table 3. Variants of the 3′ splice consensus. The top sequence is the consensus of all genes (Y = C or T).

GENE	SEQUENCE	REFERENCE
consensus	YYYYYYYYYYNYAG	
human Gs-α	TTTCAATCCCACTG	22
Drosophila Gs-α	TTTCAAATGTGCTG	23
Drosophila per	CTTCTCCTCCGCCG	24
Drosophila per	CAACGCGTTCGTCG	24

Table 4. The 5′ and 3′ end of the two known splice sites which do not show any complementarity to the 5′ end of the U1 snRNP. Gaps in each sequence denote the junction between intron and exon or vice versa.

GENE	5′ SEQUENCE	3′ SEQUENCE	REFERENCE
human proliferating cell nucleolar protein P120	AAT ATATCC	GCCCAC ATGCCC	25
chicken cartlige matrix protein	AGG ATATCC	ACTCAC TGGAAG	26

from the database, and which results in the splice being a consensus 3′ AG sequence.

d) Human myelin proteolipid gene.
Two 3′ splices in this gene are listed as non-consensus sites. When the original paper is examined (15) it can be seen that both splices have been incorrectly annotated in the database, the first splice is actually 9bp 3′ of that shown in the database, the second is 11 bp 5′.

e) Murine leukaemia virus insertion.
This sequence is from an insertion of a retrovirus into an intron of the p53 gene (16). The sequence is an interuption of an intron (i.e the point at which the retrovirus has inserted) not the 3′ end of an intron.

f) Mouse CD4 T-cell antigen.
The sequence is not of a splice site but is the junction between protein encoding and 3′ untranslated sequence (17). The database is incorrectly annotated.

g) Mouse glutathione S-transferase.
The original paper has a consensus 3′ splice site (18). The database also lists a consensus sequence, but it is 14bp 3′ of the one in the paper. The Shapiro listing has yet another site a further 8 bases 3′ (and 22 bases 3′ of the published splice)

h) Mouse *Lyt-2* T-cell differetiation antigen.
This is the end of the sequence of intron B listed in the original paper (19). Although this sequence is within an intron it is not the end of the intron. The database is incorrectly annotated.

i) Rat α-tropomyosin.
The sequence shown as a splice is not shown as such in the original paper (20), nor is it listed as one in the database. The sequence shown is at base 1176; the nearest genuine 3′ splice site is at 1146. This appears to be an error in the database search.

j) Mole rat α-crystallin.
This is not a 3′ splice, but is a 5′ splice, with the variant GC consensus, as in several other mammalian α-crystallin genes. The database FEATURES table correctly assigns the splice.

The above assessments eliminate all the non-consensus 3′ splice sites tabulated by Shapiro et al (3). I have, however, through literature searches, identified four variants which cannot be readily eliminated, and I describe these below.

VARIANT 5′ SPLICE SITES

Table 2 lists 26 5′ variant splice sites, all of which have GC in place of the 'invariant' GT at the first two positions of the intron (positions 1 and 2). Seventeen of these are from the list of Shapiro *et al* (3). One has been taken from their list of 3′ variant splice sites (the mole rat α-crystalin gene). An additional GC variant was identified in the mouse RNA polymerase II gene (which has two such variants), one has been noted in the mouse TRP-1 gene (52), three (human keratin, bovine aspartyl protease and earthworm haemoglobin c) were identified by Medline searches, and three (human and hamster APRT and mouse super oxide dismutase) were found by inspecting these sequences at the splice shown to be variant in other species. Additionally a yeast 5′ splice site with the sequence AAGA/GCATGT was identified (21), but is not considered further, although yeast genes do have the 'invariant' GT.

Table 4 shows two more splice sites, identified through Medline search, which are completely different from the prototype sequence. These are further discussed below.

VARIANT 3′ SPLICE SITES

Although all the 3′ variants identified by Shapiro *et al* have been discounted, further sites have been located by Medline search. Four of these are shown in Table 3. These are all variants of the ubiquitous AG at −2 and −1: all contain the G, but vary at −2 by having a pyrimidine. It is worth noting that all these variants are alternative splice sites, that is, each gene has a proportion of transcripts spliced at a consensus AG, and a proportion at these variants. Interestingly, although the α-subunit of the Gs proteins of human and Drosophila both have a TG 3′ site, it is in different alternative introns in either gene (22, 23). The Drosophila *per* gene apparently has two CG 3′ sites (24).

Table 4 shows two additional sites identified using Medline, which are the 3′ ends of the highly unusual 5′ splice sites.

A NOVEL AND RARE SPLICE SITE MOTIF

The two pairs of intron/exon boundary sequences shown in Table 4, from the human proliferating cell nucleolar protein P120 gene (25) and the chicken cartilage matrix protein gene (26), appear to represent members of a novel family of splice-site sequences. Neither show any significant similarity to the prototype 5′ site; but both are identical for at least the first 6 bases of the intron (ATATCC). Neither have the consensus AG at the 3′ end of the intron, but instead have the sequence CAC.

DISCUSSION

What lessons can be drawn from this study? Firstly, the sequence databases should not be regarded as error-free. As database input increases the problem of entry errors will likewise increase. Most misassignments of splice sites identified in Table 1 were, however, not errors in sequence entry, but in annotation of the sequence in the FEATURES table of the corresponding entry.

In some cases the search used by Shapiro *et al* misinterpreted correct annotations, but on the whole the FEATURES table was in error. These are not necessarily errors on the part of the database staff, but may again be faulty communication of features to the databases.

Are there any particular characteristics of the consensus sequences around the *bone fide* variant sites? There are too few sequences with variants of the 3′ AG to generalise, but all four are pyrimidine rich in the intron, as is the consensus. All four are alternative splice sites.

Almost all the 5′ variant sites have the GT dinucleotide at the first two bases of the intron replaced by GC. *In vitro* studies have shown that this substitution is the only one which will still allow the 5′ site to be accurately cleaved. albeit more slowly than the usual GT sequence (27). The rest of the sequences of the 5′ GC as a whole are significantly different from the usual GT sites. The surrounding sequences show a greater match to the prototype sequence (and therefore have greater complementarity to U1 RNA) than the average consensus splice. At base −1 (i.e. the last base of the exon) this group match the prototype G in 96% of cases, compared to 79% of all GT splices. At −2, 88% have the prototype A, compared with only 60% of GT sites. Within the intron at base +3, 80% of the GC sites have A; only 59% of GT splices have this base. At +4 the A is in 84% of GC sites and 71% of GT and at +5 the G is found in 100% of this group but in 82% of all GT splice sites.

Jacob and Gallinaro (4) have shown that mismatches between the splicing substrate and U1 RNA can be tolerated in splice sites either 5′ or 3′ of the cleavage site, but not both. The better consensus match of the GC variant splices can be explained by the pairing requirement with U1 RNA in the spliceosome, which is lacking the usual central U:A pair from the GT sequence, and so needs a better match from the rest of the sequence. This observation might be useful as an addition to the normal sequence matrix for identifying splice sites in uncharacterised DNA sequences.

I have also noted here two genes with very unusual sequences at both ends of one intron. Although the genes are completely unrelated, the sequences at the ends of the introns are strikingly similar. Both introns begin ATATCC and end PyPyCAC. It is possible that these splice sites indicate a novel (and presumably rare) splice mechanism.

ACKNOWLEDGEMENTS

Work in my laboratory is funded by the Medical Research Council and the Lister Institute for Preventive Medicine. I am a Lister Fellow.

REFERENCES

1. Mount,S.M. (1982) *Nucleic Acid. Res.*, **10**, 459–472.
2. Shapiro,M.B. and Senepathy,P. (1986) *Nucleic Acid. Res.*, **15**, 7155–7174.
3. Senepathy,P., Shapiro,M.B. and Harris,N.L. (1990) *Methods Enzymol.*, **183**, 252–278.
4. Chambon,P. and Breathnach,R. (1981) *Ann. Rev. Biochem.*, **50**, 349–383.
5. Jacob,M. and Gallinaro,H. (1989) *Nucleic Acid. Res.*, **17**, 2159–2180.
6. Ruby,S.W. and Abelson,J. (1991) *Trends Genet.*, **7**, 79–85.
7. Baeumlein,H., Pustell,J., Wobus,U., Case,S.T. and Kafatos,F.C. (1986) *J. Mol. Evol.*, **24**, 72–82.
8. Zucker,C.S., Cowman,A.F. and Rubin,G.M. (1985) *Cell*, **40**, 851–858.
9. Boss,J.M., Mengler,R., Okada,K., Auffray,C. and Strominger,J.L. (1985) *Mol. Cell. Biol.*, **5**, 2677–2683.
10. Selby,M.J., Edwards,R., Sharp,F. and Rutter,W.J. (1987) *Mol. Cell. Biol.*, **7**, 3057–3064.
11. Baron,A., Featherstone,M.S., Hill,R.E., Hall,A., Galliot,B. and Duboule,D. (1987) *EMBO J.*, **6**, 2977–2986.
12. Schneuwly,S., Kuroiwa,A., Baumgartner,P. and Gehring,W.J. (1986) *EMBO J.*, **5**, 733–739.
13. Roberts,R.J., O'Neill,K.E. and Yen,C.T. (1984) *J. Biol. Chem.*, **259**, 13968–13975.
14. Kropp,K., Gulick,J. and Robbins,J. (1986) *J. Biol. Chem.*, **261**, 6613–6618.
15. Diehl,H.-J., Schaich,M., Budzinski,R.-M. and Stoffel,W. (1986) *Proc. Nat. Acad. Sci. U.S.A.*, **83**, 9807–9811.
16. Wolf,D. and Rotter,V. (1984) *Mol. Cell. Biol.*, **4**, 1402–1410.
17. Gorman,S.D., Tourvieille,B. and Parnes,J.R. (1987) *Proc. Nat. Acad. Sci. U.S.A.*, **84**, 7644–7648.
18. Daniel,V., Sharon,R., Tichauer,Y. and Sarid,S. (1987) *DNA*, **6**, 317–324.
19. Liaw,C.W., Zamoyska,R. and Parnes,J.R. (1986) *J. Immunol.*, **137**, 1037–1043.
20. Ruiz-Opazo,N. and Nadel-Ginard,B. (1987) *J. Biol. Chem.*, **262**, 4755–4765.
21. Hodge,M.R. and Cumsky,M.G. (1989) *Mol. Cell. Biol.*, **9**, 2765–2770.
22. Kozasa,T., Itoh,H., Tsukamoto,T. and Kaziro,Y. (1988) *Proc. Nat. Acad. Sci. U.S.A.*, **85**, 2081–2085.
23. Quan,F. and Forte,M.A. (1990) *Mol. Cell. Biol.*, **10**, 910–917.
24. Citri,Y., Colot,H.V., Jacquier,A.C., Yu,Q., Hall,J.C., Baltimore,D. and Rosbach,M. (1987) *Nature*, **326**, 42–47.
25. Larson,R.G., Henning,D., Haidar,M.A., Jhiang,S., Lin,W.L., Zhang,W.W. and Busch,H. (1990) *Cancer Commun.*, **2**, 63–71.
26. Kiss,I., Deak,F., Holloway,R.G., Delius,H., Mebus,K.A., Frimberger,E., Argraves,W.S., Tsonis,P.A., Winterbottom,N. and Goetinck,P.F. (1989) *J. Biol. Chem.*, **264**, 8126–8134.
27. Aebi,M., Hornig,H. and Weissmann,C. (1987) *Cell*, **50**, 237–246.
28. Shibahara,S., Kubo,T., Perski,H.J., Takehashi,H., Noda,M. and Numa,S. (1985) *Eur. J. Biochem*, **146**, 15–22.
29. Cool,D.E. and MacGillivray,R.T.A. (1987) *J. Biol. Chem.*, **262**, 13662–13673.
30. Hagen,F.S., Gray,C.L., O'Hara,P., Grant,F.J., Saari,G.C., Woodbury,R.G., Hart, C.E., Insley,M., Kisiel,W., Kurachi,K. and Davies,E.W. (1986) *Proc. Nat. Acad. Sci. U.S.A.*, **83**, 2412–2418.
31. Morohashi,K., Sogawa,K., Omura,T. and Fujii-Kuriyama,Y. (1987) *J. Biochem.*, **101**, 879–887.
32. Degen,S.J.F. and Davie,E.W. (1986) *Biochemistry*, **26**, 6165–6177.
33. Hikada,Y., Tarle,S.A., O'Toole,T.E., Kelley,W.N. and Palella,T.D. (1987) *Nucleic Acid. Res.*, **15**, 9086.
34. Kulesh,D.A. and Oshima,R.G. (1989) *Genomics*, **4**, 339–347.
35. Levanon,D., Lieman-Hurwitz,J., Wigderson,M., Sherman,L., Bernstein,Y., Laver-Rudich,Z., Daneiger,E., Stein,O. and Groner,Y. (1985) *EMBO J.*, **4**, 77–84.
36. Bendetto,M.T., Anzai,Y. and Gordon,J.W. (1991) *Gene*, **99**, 191–195.
37. van den Heuval,R., Hendricks,W., Quax,W. and Bloemendal,H. (1985) *J. Mol. Biol.*, **185**, 273–284.
38. King,C.R. and Piatigorsky,J. (1986) *Cell*, **32**, 707–712.
39. Hendricks,W., Leunissen,J., Nevo,E., Bloemendal,H. and de Jong,W.W. (1987) *Proc. Nat. Acad. Sci. U.S.A.*, **84**, 5320–5324.
40. Nalbantoglu,J., Phear,G.A. and Meuth,M. (1986) *Nucleic Acid. Res.*, **14**, 1914.
41. Dush,M.K., Sikela,J.M., Khan,S.A., Tischfield,J.A. and Stambrook,P.J. (1985) *Proc. Nat. Acad. Sci. U.S.A.*, **82**, 2731–2735.
42. Ahearn,J.M., Bartolomei,M.S., West,M.L., Cisek,L.J. and Corden,J.L. (1987) *J. Biol. Chem.*, **262**, 10695–10705.
43. Mueller,R.M., Taguchi,H. and Shibahara,S. (1987) *J. Biol. Chem.*, **262**, 6795–6802.
44. Vize,P.D. and Wells,J.R.E. (1987) *Gene*, **55**, 339–344.
45. Lu,Q., Wolfe,K.H. and McConnell,D.J. (1988) *Gene*, **71**, 135–146.
46. Molina,M.I., Kropp,K.E., Gulick,J. and Robbins,J. (1987) *J. Biol. Chem.*, **262**, 6473–6488.
47. Dodgson,J.B. and Engel,J.D. (1983) *J. Biol. Chem.*, **258**, 4623–4629.
48. Erbil,C. and Niessing,J. (1983) *EMBO J.*, **2**, 1339–1343.
49. Jhiang,S.M. and Riggs,A.F. (1989) *J. Biol. Chem.*, **264**, 19003–19008.
50. Huiet,L. and Giles,N.H. (1986) *Proc. Nat. Acad. Sci. U.S.A.*, **83**, 3381–3385.
51. Katinakis,P. and Verma,D.P.S. (1985) *Proc. Nat. Acad. Sci. U.S.A.*, **82**, 4157–4161.
52. Jackson, I.J., Chambers, D.M., Budd, P.S. and Johnson, R. (1991) *Nucleic Acids Res.* **19**, 3799–3804.