RNA recognition by the human immunodeficiency virus Tat and Rev proteins

Michael J. Gait and Jonathan Karn

The human immunodeficiency virus (HIV-1) regulatory proteins, Tat and Rev, are important potential targets for the development of new drug therapies against HIV infection. Both proteins are highly specific RNA-binding proteins that recognize *cis*-acting regulatory elements in the viral mRNAs. These interactions are fascinating paradigms of a new principle of RNA recognition in which the protein makes contact with functional groups displayed in a distorted major groove of an RNA duplex.

IN THE SEARCH for new therapeutic targets within the human immunodeficiency virus (HIV) – the 'Achilles heel' against which effective drugs may be developed – much recent attention has focused on the viral regulatory proteins Tat and Rev. These two proteins are unique to HIV and closely related lentiviruses, where they play complementary roles in the virus life cycle. Tat stimulates transcription from the viral long terminal repeat (LTR), whereas Rev is required for the efficient expression of the late mRNAs encoding the structural proteins of the virus.

Tat and Rev both achieve their effects by binding to *cis*-acting regulatory elements encoded by the viral mRNAs; All other retroviral regulatory proteins studied previously interact with DNA sequences. Studies of the interactions of Tat and Rev with RNA are providing new insights into the chemistry of nucleic acid recognition. There are several examples of RNA binding proteins that recognize bases displayed in apical loop and bulge structures[1]. By contrast, both Tat and Rev interact with functional groups on Watson–Crick base pairs that are exposed within the major groove of a distorted double-stranded RNA. This new principle of RNA recognition is likely to extend to many other protein–RNA interactions.

Regulatory circuits

Several recent reviews have described the genetic organization of HIV[2–4]. To summarize briefly, Tat activity requires the *trans*-activation-responsive region (TAR), a stem–loop structure found at the 5' terminus of all the mRNAs transcripts (Fig. 1). Recent evidence suggests that Tat acts largely to stimulate elongation by RNA polymerase II, rather than the initiation of transcription. In the absence of Tat, most of the RNA polymerases engaged in transcription stall near the 3' end of the TAR RNA stem. Tat relieves this transcriptional polarity and stabilizes RNA polymerases downstream of the promoter.

Transcription of the HIV genome is also characterized by a progressive shift from the synthesis of short, multiply spliced mRNAs, which encode the viral regulatory proteins, towards the production of late mRNAs which encode the virion proteins. Several distinct mechanisms of action have been suggested for Rev, including the coupling of Rev activity to splicing itself, but it now seems most likely that Rev acts either to stimulate the export of partially spliced viral mRNAs from the nucleus, or to increase the stability of viral mRNAs in the cytoplasm of infected cells. Rev activity requires binding to the Rev-response element (RRE), a region of extensive RNA stem–loops located within the coding sequence of the *env* gene (Fig. 1).

Tat binding to the TAR RNA 'bulge'

Highly purified recombinant Tat expressed in *Escherichia coli* forms one-to-one complexes with TAR RNA with $K_d \approx$ 3 nM (Refs 5–7). The binding site for Tat on TAR RNA is defined by a U-rich trinucleotide bulge in the upper stem (Fig. 2). Essential residues for Tat recognition are U23 and the two base pairs immediately above the bulge, G26:C39 and A27:U38[6–9]. The other residues in the bulge, U or C24 and U25, appear to act predominantly as spacers and may be replaced by other nucleotides, or even by non-nucleotide linkers[10]. The two base pairs below the bulge make only a small contribution to Tat binding specificity and are probably not points of direct contact with the protein. Similarly, sequences in the apical loop of TAR RNA are not required for Tat binding even though this region of the RNA is essential for efficient *trans*-activation[5,6,9]. A likely role for the loop is to act as a binding site for cellular cofactors of Tat.

Recognition of specific bases in the major groove

What can be said about the structural basis for the Tat–TAR interaction? Gel retardation studies have shown that the bulge causes bending of the RNA duplex[11]. Weeks and Crothers[12] have postulated that the U-rich bulge also creates a distortion that opens the otherwise narrow major groove of the TAR RNA duplex. This important insight was based on the observation that both G26 and A27 show increased reactivity to diethylpyrocarbonate (DEPC) when a bulge of three uridine residues was present[12]. Functional group mutagenesis studies have also provided strong evidence in support of the hypothesis[13]: Tat binding is considerably reduced by methylation of the O^4 or N^3-H of U23 or by removal of either of the N^7 nitrogen atoms of G26 or A27 from the major groove[10,13]. By contrast, removal of the exocyclic amino group of G26 from the minor groove does not affect Tat binding[13]. Thus, hydrogen bonding in the major groove of the RNA duplex is essential for Tat binding.

Phosphate contacts

Precisely how many Tat–TAR contacts are base-specific and how many involve phosphate and sugar contacts remains unclear[7]. Tat has a very low affinity for RNAs that lack a suitable bulge structure. For example, Tat–TAR binding is 1000-fold stronger that Tat bound to tRNA. However, the variation in the binding energies between wild-type TAR RNA and the TAR RNA mutants that do not respond to Tat *in*

M. J. Gait and **J. Karn** are at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 2QH.

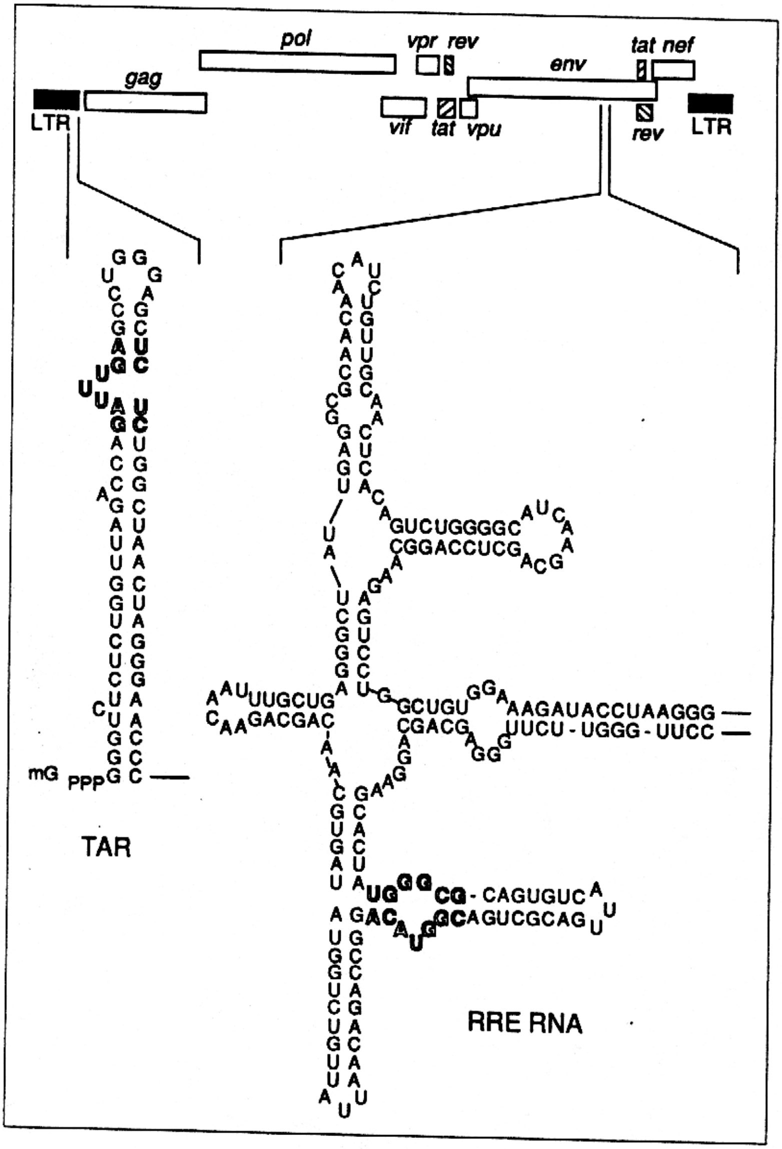

Figure 1
Location of the TAR and RRE elements in the HIV genome and their proposed secondary structures. Highlighted residues are those which are implicated in specific recognition by the Tat and Rev proteins, respectively.

vivo is comparatively small – only 1.25 kcal mol^{-1} or approximately the energy assigned to one hydrogen bond. This suggests that base-specific interactions between Tat and TAR RNA contribute only a small amount of binding energy on top of the high overall affinity of Tat for an RNA with the correct bulge structure. This view is supported by recent mapping of the phosphate contacts between Tat and TAR RNA by chemical-modification interference experiments[7]. Ethylation of P23 or any one of five phosphates (P36–P40) located on the strand opposite the U-rich bulge caused substantial interference with Tat binding. Similarly, removal of the negative charge at the phosphate between A22 and U23 (P23) by methylphosphonate incorporation, also strongly interferes with Tat binding[13].

Both 'basic' and 'core' regions required for specific binding

Which residues in the Tat protein are important for binding to TAR RNA? Notable features of the 86-residue Tat protein from HIV-1 include a cysteine-rich region (residues 22–31) and a 'core' sequence (residues 32–47) which are highly conserved between immunodeficiency viruses. The cysteines allow Tat to bind two molecules of zinc[6,14], but the suggestion that Tat exists as a zinc-bound dimer[15] is now discounted[6,16]. Towards the carboxyl terminus of Tat is a highly basic region (residues 48–57) which appears to be involved in RNA binding as well as in directing Tat to the nucleus.

Purification of Tat from genetically engineered *E. coli* is difficult, since its seven cysteines oxidize readily and Tat tends to associate with bacterial RNA. To circumvent these problems, several groups used peptides as models of the Tat protein. Initially this approach looked quite promising. Weeks *et al.*[17] found that a proteolytic fragment of Tat carrying the basic region and carboxyl terminus was able to bind to TAR RNA. Even shorter basic peptides are able to bind to TAR RNA with affinities approaching that of Tat[18,19], and like the protein, the peptides bind selectively to molecules carrying a bulge and the G26 and A27 bases in the stem[12,18–20]. Unfortunately the binding specificities of basic peptides for TAR RNA only superficially resemble that of the Tat protein. Recent detailed comparisons of peptide and protein binding have shown that there are important differences which were originally overlooked[7,20].

When peptide and protein binding were compared using a sensitive dual-label filter-binding assay and refined gel mobility shift assays, it was found that a short peptide spanning the basic region (residues 48–72) failed to exhibit much discrimination between wild-type and mutant TAR sequences[7,20]. By contrast, a longer peptide (residues 37–72) containing additional residues from the conserved 'core' region had a higher affinity for TAR RNA and was able to discriminate between TAR RNA mutants with a specificity that closely resembled that of the Tat protein[7]. These binding differences are also apparent in comparisons between the DEPC and ethylation interference patterns for peptide and protein binding. Peptides with reduced specificity for TAR RNA show reduced numbers of contacts with the bases and backbone phosphates[7].

What role could the core sequence play in TAR recognition? The data suggest that small basic Tat peptides probably do not make the same contacts as the Tat protein, even though both the protein and the peptides appear to align in the major groove of TAR RNA. One possibility is that the core region contributes to Tat binding specificity by orienting the basic region within the major groove. Alternatively, it is possible that critical amino acid residues in the core region, such as Lys41, could form base-specific contacts in the major groove of TAR while the basic region forms electrostatic contacts with the surrounding phosphates.

How does the basic region of Tat participate in TAR RNA recognition? Many RNA binding proteins, including the bacteriophage λ N protein[21], carry arginine-rich regions which appear to participate in RNA binding. On the basis of peptide experiments, Frankel and colleagues proposed that a single arginine is all that is required for specific TAR recognition[18]. In their original hypothesis[18] called the 'arginine-fork', it was proposed that the arginine formed bifurcated hydrogen bonds with only two phosphate residues in TAR (P22 and P23). Nuclear magnetic resonance studies of a complex between the amino acid derivative argininamide and the top 31 residues of TAR RNA have led Puglisi *et al.*[22] to a refinement of the 'arginine-fork' model. The NMR data suggest that in the presence of the modified amino acid a base triple is formed between U23 and A27:U38. This creates a binding pocket which allows the argininamide to insert into the major groove by forming two hydrogen bonds with G26 and electrostatic contacts with P22 and P23. To achieve this structure, the TAR RNA is dramatically rearranged, with the U-rich bulge forced out of the RNA duplex.

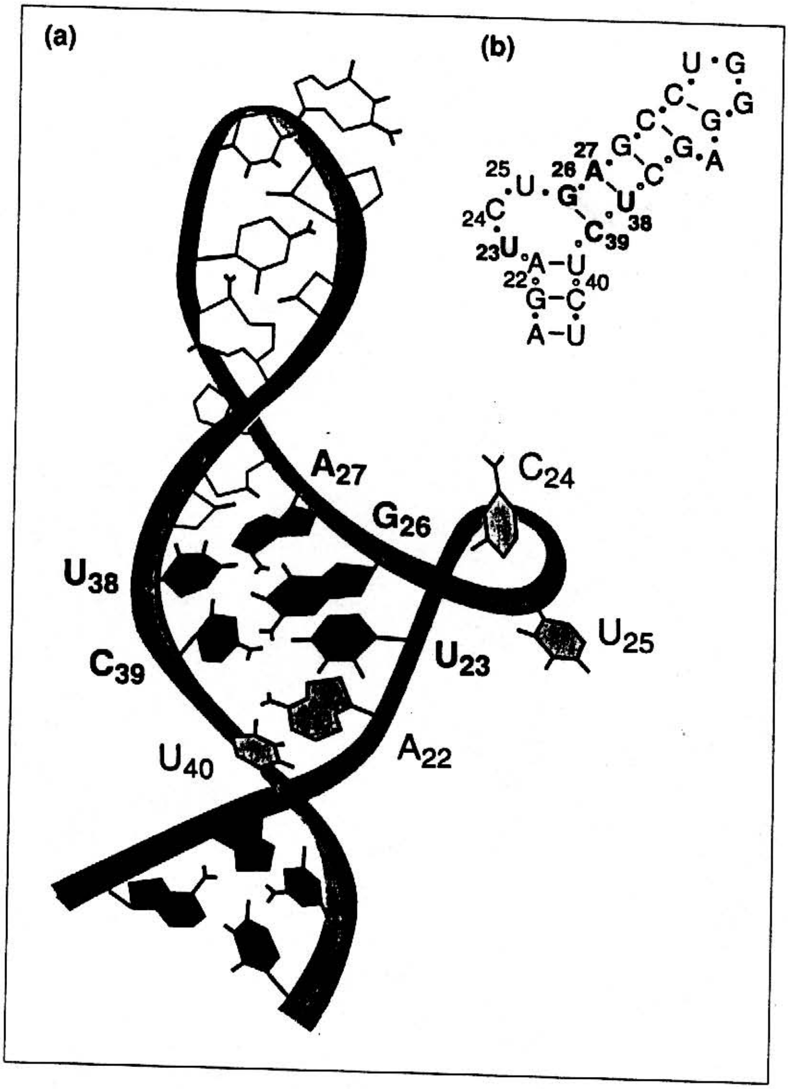

Figure 2
(a) Three-dimensional representation of the Tat-binding site on TAR RNA. The structure is based on a model proposed by Loret *et al.*[23] and is similar to that proposed by Weeks and Crothers[20] and the model for free TAR described by Puglisi *et al.*[22] Residues in black are those implicated in direct recognition of Tat whereas those in grey are presumed to play a largely structural role. **(b)** Secondary structure of the apical region of TAR. Open circles indicate the positions of phosphates, where ethylation interferes with Tat binding[13].

Although the data for the TAR–argininamide structure is strong, we (the authors) are still unconvinced that this interaction obeys the same rules as Tat–TAR recognition. Tat binding is achieved at nanomolar concentrations, whereas argininamide binds only at millimolar concentration, a difference in binding constant of 10^6. Additional contact between Tat and RNA must be responsible for this enormous difference. Base-substitution experiments have shown that the same base residues are required both for argininamide binding and for Tat recognition[7,8,18], but such experiments cannot determine whether the type of recognition is identical. However, one of our recent functional group mutagenesis experiments suggests that the specificities of argininamide and protein binding may be different. The Tat protein binds with high affinity to TAR RNA carrying N^6-methyl-A27, a modified base that would be unable simultaneously to form a base triple with U23 and make a contact with Tat[13]. The second central feature of the Puglisi *et al.* model is that the RNA is rearranged upon argininamide binding. It is unclear whether Tat induces a similar conformational shift in TAR RNA since Tat-derived peptides do not produce a substantial change in the CD spectrum of TAR RNA upon binding[23]. We will have to await a full structure determination of the RNA–protein complex for the definitive answer as to how Tat recognizes TAR.

RNA recognition by Rev

Rev from HIV-1 is 116 residues long and, like Tat, it also carries an arginine-rich sequence (residues 34–50) that participates in both RNA binding and nuclear localization. Indeed, the two arginine-rich sequences can be exchanged without inhibiting the activity of either protein[24]. The RNA region required for Rev activity is remarkably

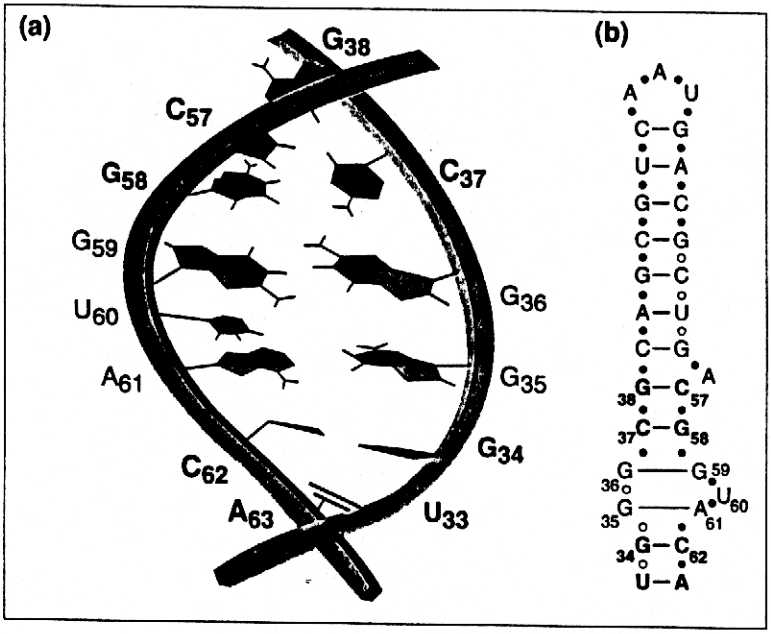

Figure 3

(a) Three-dimensional representation of the high affinity Rev binding site on the RRE RNA. Residues in grey are thought to play a largely structural role whereas those in black are implicated in recognition by Rev. U60 has been arbitrarily modelled stacked into the helix. **(b)** Secondary structure of the region of RRE RNA involved with high-affinity Rev binding. Open circles indicate the positions of phosphates where ethylation interferes with Rev binding[38].

large in comparison to the interaction site for Tat on TAR RNA. The RRE RNA as defined by deletion analysis is more than 300 nucleotides long and has an elaborate secondary structure (Fig. 1)[25]. Virtually the entire RRE sequence is required for the full biological activity of Rev *in vivo*[25].

RNA transcripts containing the RRE are specifically bound by Rev *in vitro*[26–28]. The binding reaction is complex and involves an initial interaction with a high-affinity site characterized by an unusual 'bubble' structure[29] (Fig. 3). Subsequently, additional Rev molecules bind in a cooperative manner to lower affinity sites on the flanking RNA sequences[28]. This oligomerization reaction results in the production of a series of discrete complexes that can be separated by non-denaturing gel electrophoresis[28–31]. The regular nature of the secondary structure for the RRE (Fig. 1) suggests that, after initial recognition of the bubble on one arm, there is an ordered recognition of the other five arms. The precise details of Rev assembly on the RRE are not yet understood, although it is clear that between six and eight Rev monomers bind to the 234 nt-long RRE RNA[27,29]. At high protein concentrations additional Rev monomers are added and the RRE RNA is packaged into rod-like ribonucleoprotein filaments[29,32].

Mutagenesis studies strongly suggest that after nucleation at the high-affinity Rev binding site, oligomerization at flanking sites takes place *in vivo*. Rev activity is abolished by deletion of the high-affinity binding site, as well as by the introduction of mutations into the bubble that block high affinity binding *in vitro*[33]. Similarly, mutations that disrupt double-stranded regions in the RRE invariably inactivate Rev function *in vivo*[34–36].

High affinity binding by Rev at the 'bubble' structure

Recognition of the high-affinity binding site by Rev shows strong parallels to the recognition of TAR RNA by Tat. Both proteins recognize base pairs in a distorted major groove of duplex RNA, as well as making contacts with the adjacent phosphates.

Synthetic RNA duplexes containing the bubble structure and flanking base pairs, are able to form a one-to-one complex with Rev (K_d ~ 3 nM; Refs 37, 38). The two base pairs on either side of the bubble are involved in base-specific contacts with the protein. N^7-Carboxyethylation by DEPC of any of the purines flanking the bulged bases[38,39], or removal of N^7 from G34[37], blocks Rev binding. However, the bubble sequence is not sufficient for Rev binding. Short RNA duplexes with less than six flanking base pairs are unable to bind Rev with high affinity, suggesting that the region of duplex RNA covered by Rev must extend beyond the bubble structure[37]. Ethylation interference experiments have shown that Rev makes contact with phosphates located on both strands of the duplex (P33–35 and P52–54)[38]. Similarly, footprinting data show that 10–14 base pairs are covered by a Rev molecule that is positioned asymmetrically with respect to the bubble[38].

Non-Watson–Crick base pairs stabilize the bubble

The nature of the distortion in the bubble is clearly different from that introduced by the bulge in TAR RNA (Fig. 3). The bubble structure is believed to be stabilized by non-Watson–Crick base pairs[29,33] formed between the bulged residues. Functional group mutagenesis of G residues in the bubble has shown that the exocyclic groups of all three of G35, G36 and G59 can be removed without loss of Rev binding[37]. Thus the G36 : G59 pair must form hydrogen bonds via their respective O^6 and N^1-H positions. The structure of the G35 : A61 pair is not yet known; two possible pairing schemes are consistent with the data. The only truly single-stranded residue in the bubble appears to be U60, the identity of which seems to be unimportant for Rev binding[29,33]. Model building suggests that there is relatively little bending at the bubble, but the non-Watson–Crick pairs plus the presence of the U60 residue allows sufficient opening of the major groove for specific recognition to take place at G34 and other flanking base residues.

Low affinity binding by Rev

It seems likely that the low affinity interactions of Rev with RNA sequences flanking the bubble are very similar to the high-affinity binding interactions. Because the major groove is narrow in the regular duplex RNA, Rev is unlikely to form contacts with the bases, and phosphate contacts, probably involving the arginines, are expected to contribute much of the binding energy.

Protein–protein contacts are also used to stabilize Rev binding to the lower affinity sites. The aggregation properties of the Rev protein even allow filament formation in the absence of RNA[29,32]. Mutations in the regions on either side of the arginine-rich region (residues 20–33 and residues 51–65) cause defects in formation of multiple complexes with RRE RNA[31]. A detailed analysis of these 'multimerization' domains should reveal much about the packing of Rev on the RNA.

Biological implications

How relevant are the *in vitro* measurements of Tat and Rev binding to the physiological situation? For both proteins there is a strict correlation between their ability specifically to recognize RNA sequences and the efficiency of the *in vivo* response. The RNA-binding properties of the two proteins also make it likely that the distinct kinetic phases of HIV transcription are a reflection of their intracellular levels. Assuming a diameter of 5–10 μm for the nucleus of a typical human cell, the presence of only 40–320 molecules of Tat in the nucleus would correspond to a concentration of 1 nM in the absence of competitor RNAs. At these concentrations, approximately 30% of the nascent transcripts from an integrated provirus would be expected to be bound by Tat[7]. Since Tat is functional as a monomer and has a catalytic effect on transcription, this calculation suggests how a strong stimulation of transcription can be achieved early during the virus growth cycle. As the rate of transcription increases, the levels of Rev rise. Once Rev concentrations are above 10 nM oligomerization of Rev on the RRE is expected to occur and late mRNA production can begin. Of course, the *in vivo* situation is likely to be more complicated than these simple calculations suggest. Additional proteins may participate in the formation of complexes between Tat, TAR and the RNA polymerase that may help to further stabilize the Tat–TAR interaction.

Conclusions and future challenges

The detailed biochemical studies of the HIV regulatory and maturation proteins described above have revealed a new principle for RNA recognition: both proteins make specific contacts at Watson–Crick base pairs which become accessible because of the structural perturbations introduced by bulged bases or by non-Watson–Crick base pairs. In addition to these base-specific contacts, both proteins make phosphate contacts on both strands of the flanking RNA duplexes. These observations raise the exciting possibility that small molecules competing for functional groups in the binding sites on the RNA might be able to inhibit Tat and Rev binding, and thus their biological activity. Biochemical assays for RNA binding and *trans*-activation could easily be adapted for the large-scale screening of compounds with inhibitory activity. There may even be some prospects for specific drug design, since the three-dimensional structures of the binding sites on TAR and RRE RNA should soon be known.

Acknowledgements

We thank Clare Pritchard and Steve Salisbury for help with molecular modelling of the Rev binding site and the many members of the Laboratory of Molecular Biology involved with the study of the HIV Tat and Rev proteins for discussions and comments on the manuscript.

References

1 Nagai, K. (1992) *Curr. Opinion Struct. Biol.* 2, 131–137
2 Kam, J. (1991) *Curr. Opinion Immunol.* 3, 526–536
3 Cullen, B. R. and Malim, M. H. (1991) *Trends Biochem. Sci.* 16, 346–350
4 Rosen, C. A. (1991) *Trends Genet.* 7, 9–14
5 Dingwall, C. *et al.* (1989) *Proc. Natl Acad. Sci. USA* 86, 6925–6929
6 Dingwall, C. *et al.* (1990) *EMBO J.* 9, 4145–4153
7 Churcher, M. *et al.* (1993) *J. Mol. Biol.* 230, 90–110
8 Delling, U. *et al.* (1992) *J. Virol.* 66, 3018–3025
9 Roy, S. *et al.* (1990) *Genes Dev.* 4, 1365–1373
10 Sumner-Smith, M. *et al.* (1991) *J. Virol.* 65, 5196–5202
11 Riordan, F. A., Bhattacharyya, A., McAteer, S. and Lilley, D. M. J. (1992) *J. Mol. Biol.* 226, 305–310
12 Weeks, K. M. and Crothers, D. M. (1991) *Cell* 66, 577–588
13 Hamy, F. *et al.* (1993) *J. Mol. Biol.* 230, 111–123
14 Slice, L. W. *et al.* (1992) *Biochemistry* 31, 12062–12068
15 Frankel, A. D., Bredt, D. S. and Pabo, C. O. (1988) *Science* 240, 70–73
16 Rice, A. P. and Chan, F. (1991) *Virology* 185, 451–454
17 Weeks, K. M. *et al.* (1990) *Science* 249, 1281–1285
18 Calnan, B. J. *et al.* (1991) *Science* 252, 1167–1171
19 Cordingley, M. G. *et al.* (1990) *Proc. Natl Acad. Sci. USA* 87, 8985–8989
20 Weeks, K. M. and Crothers, D. M. (1992) *Biochemistry* 31, 10281–10287
21 Lazinski, D., Grzadzielska, E. and Das, A. (1989) *Cell* 59, 207–218
22 Puglisi, J. D. *et al.* (1992) *Science* 257, 76–80
23 Loret, E. P., Georgel, P., Johnson, W. C. J. and Ho, P. S. (1992) *Proc. Natl Acad. Sci. USA* 89, 9734–9738
24 Subramanian, T., Govindarajan, R. and Chinnadurai, G. (1991) *EMBO J.* 10, 2311–2318
25 Malim, M. H. *et al.* (1989) *Nature* 338, 254–257
26 Zapp, M. L. and Green M. R. (1989) *Nature* 342, 714–716
27 Daly, T. J. *et al.* (1989) *Nature* 342, 816–819
28 Heaphy, S. *et al.* (1990) *Cell* 60, 685–693
29 Heaphy, S. *et al.* (1991) *Proc. Natl Acad. Sci. USA* 88, 7366–7370
30 Kjems, J., Brown, M., Chang, D. D. and Sharp, P. A. (1991) *Proc. Natl Acad. Sci. USA* 88, 683–687
31 Malim, M. H. and Cullen, B. R. (1991) *Cell* 65, 241–248
32 Wingfield, P. T. *et al.* (1991) *Biochemistry* 30, 7527–7534
33 Bartel, D. P., Zapp, M. L., Green, M. R. and Szostak, J. W. (1991) *Cell* 67, 529–536
34 Cochrane, A. W., Chen, C-H. and Rosen, C. A. (1990) *Proc. Natl Acad. Sci. USA* 87, 1198–1202
35 Holland, S. M., Chavez, M., Gerstberger, S. and Venkatesan, S. (1992) *J. Virol.* 66, 3699–3706
36 Dayton, E. T. *et al.* (1992) *J. Virol.* 66, 1139–1151
37 Iwai, S. *et al.* (1992) *Nucleic Acids Res.* 20, 6465–6472
38 Kjems, J., Calnan, B., Frankel, A. D. and Sharp, P. A. (1992) *EMBO J.* 11, 1119–1129
39 Tiley, L. S. *et al.* (1992) *Proc. Natl Acad. Sci. USA* 89, 758–762

TIBS reference lists

Authors of *TIBS* articles are asked to limit the number of references cited to provide non-specialist readers with a concise list for further reading. It is hoped that the citation of other, more extensive review articles rather than a comprehensive list of original articles enables interested readers to delve more immediately into the topic.