OPINION

The principles of guiding by RNA: chimeric RNA–protein enzymes

Alexander Hüttenhofer and Peter Schattner

Abstract | The non-protein-coding transcriptional output of the cell is far greater than previously thought. Although the functions, if any, of the vast majority of these RNA transcripts remain elusive, out of those for which functions have already been established, most act as RNA guides for protein enzymes. Common features of these RNAs provide clues about the evolutionary constraints that led to the development of RNA-guided proteins and the specific biological environments in which target specificity and diversity are most crucial to the cell.

The early view of non-protein-coding RNAs (ncRNAs) was that they are relics of a primordial 'RNA world' in which RNA served both as the carrier of genetic information and as the catalytic agent. The current view of the RNA world is far more complex. True catalytic RNAs (ribozymes) are in fact rare. Instead, most ncRNAs carry out their cellular duties by various mechanisms that are not directly catalytic (see TABLE 1 for some common ncRNA species and their functions). A few RNAs, such as SRP RNA (signal recognition particle RNA)[1,2], seem to function as obligate cofactors of catalytic protein complexes. Some ncRNAs, such as 7SK (REF. 3), 6S RNA[4], *CsrB* and *CsrC*[5] and perhaps *Air* RNA[6], function as genetic regulators by means of antagonistic competition for protein binding sites. Others have a structural role or function as scaffolds onto which catalytic proteins can assemble. Therefore, the numerous ncRNAs might be classified according to their functions, such as: catalysts, guides, catalytic cofactors, antisense RNAs, protein binding-site antagonists/ agonists or templates (TABLE 2). This classification has its limitations: many RNAs (for example, *Air*, *H19*) function in ways that are still not completely understood, and some RNAs have more than one function. For example, SRP RNA is both a structural scaffold and a cofactor[1,2], and telomerase RNA, although primarily a template for the repetitive polymerization of the telomere end sequence, might also assist in the guiding of the telomerase ribonucleo-protein (RNP) complex to the telomere[7].

Despite this variety of mechanisms, by far the largest number of characterized ncRNAs function as some form of guide RNA in which an RNA is used to guide or target an RNP complex to a nucleic-acid sequence (TABLE 3). RNAs that function this way include the small interfering RNAs (siRNAs), microRNAs (miRNAs) and small nucleolar RNPs (snoRNAs), and mediate a wide range of cellular functions. Here we turn to this important class of ncRNAs in search of clues about the conditions that cause nature to assign some of the workload to RNAs rather than to proteins.

RNA rules

Guide RNAs function as part of a catalytic RNP complex in which the RNA carries out the task of substrate recognition and a protein component carries out catalysis (FIG. 1a). Because the two essential components belong to different classes of macromolecules we refer to these RNA-guided proteins as 'chimeric RNP enzymes'. In general, chimeric RNP enzymes contain an unvarying protein-based enzyme portion (consisting of one or more proteins) that associates with different small guide RNAs that target the complex to its substrate by antisense complementarity (FIG. 1b).

The range of catalytic 'payloads' that are guided by these RNAs is strikingly wide and includes endonucleases, polymerases and DNA, RNA and histone methyltransferases (TABLE 3). Guide RNAs belong to a few large families, three of which — siRNA/miRNAs, snoRNAs and classical guide RNAs (gRNAs) — contain hundreds of representatives.

siRNAs/miRNAs. The siRNA–ribonucleoprotein complexes (siRNPs) target mRNAs for degradation[8]. The RNA moiety of the siRNP — a 21 nt long RNA molecule that is known as siRNA — associates with a protein complex, the RNA induced silencing complex (RISC), which cleaves mRNAs at the site of complementarity to the respective siRNA[9]. The protein content of RISC has been determined by isolating functional complexes of varying protein composition that were able to cleave mRNAs. Minimal active RISCs of about 150 kD might contain only so-called argonaute proteins (AGO) that are associated with the siRNA guide[10]. So far, four AGO proteins have been isolated in mammals; AGO2 — a protein with an RNase H-like cleavage domain — has recently been identified as the protein that is responsible for mRNA cleavage[11].

In plants, miRNAs, which are also approximately 21 nt in size, bring about mRNA cleavage by a mechanism that is identical to that of siRNAs[12]. By contrast, mammalian miRNAs are thought to function primarily by binding to 3′ UTRs of target mRNAs, thereby repressing their translation or directing mRNA destabilization by a mechanism that is different from AGO-mediated cleavage[12]. Although it has not been directly demonstrated which part of the miRNP is responsible for this function, it is believed that, similar to siRNP, miRNA targets the miRISC to the 3′ UTR of the mRNA, which then exerts its function through the AGO protein[13]. Therefore, target recognition of the mRNA by the RNA–enzyme complex, through the antisense ncRNA, is restricted to the 21 nt long antisense element, whereas the enzymatic function (mRNA cleavage or translational repression) is exerted by (a) specific associated protein(s). Recent data indicate that

Table 1 | **Common ncRNAs and their functions**

ncRNA type	Description	Function
RNase P	~400 nt long	Cleaves tRNA precursors to result in mature 5′ ends; as a catalytic RNA (ribozyme) in bacteria, for example, cleaves tRNA precursors under high monovalent salt conditions in the absence of a protein
miRNA	Small, 21–23 nt long ssRNA	Targets mRNAs for cleavage (in plants) or translation inhibition (in mammals)
siRNA	21–23 nt long ssRNA	Targets mRNAs for cleavage
rasiRNA	Small, 21–23 nt long ssRNA	Involved in repeat silencing
snoRNA	~50–200 nt long, structured RNA that is localized to the nucleolus	Specifies modification of rRNAs, snRNAs or tRNAs (in Archaea only); C/D box snoRNAs specify 2′-O-methylation of the ribose of a target RNA, H/ACA box snoRNAs specify pseudouridylation
gRNA	Small, ~60 nt long ssRNA, containing a poly U tract at its 3′ end (from 5–20 U residues)	Guides U insertions or deletions within mitochondrial pre-mRNAs of certain protozoan organisms, for example, trypanosomes
snRNA	Structured; ~100–300 nt long (in humans)	Guides splicing of pre-mRNAs (for example, U1, U2, U4, U5 and U6 snRNAs)
rRNA	Highly structured; sized between ~120 (5S rRNA) and several thousand nucleotides (18S, 28S rRNAs)	As part of the ribosome it catalyses peptide bond formation (for large rRNA only)
Xist RNA	~17 kb long RNA, which is transcribed from the X chromosome	Involved in X-chromosome inactivation and dosage compensation
tRNA	Highly structured, sized between ~70 and 95 nt	RNA adapter molecules for amino acids; guides amino acids to the ribosome in an mRNA-dependent mode
SRP RNA	Has a rod-like structure, sized ~300 nt (in humans)	Part of the SRP, a ribonucleo-protein complex that is involved in targeting specific proteins to the endoplasmic reticulum for subsequent secretion
6S RNA	Highly structured RNA (~180 nt long in *Escherichia coli*), which forms a single hairpin that is found in bacteria	Binds to the σ^{70} factor of the RNA polymerase complex, thereby regulating transcription of σ^{70} promoters

gRNA, classical guide RNA; miRNA, microRNA; ncRNA, non-protein-coding RNA; rasiRNA, repeat-associated siRNA; rRNA, ribosomal RNA; siRNA, small interfering RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA; SRP, signal recognition particle; Xist, X-(inactive)-specific transcript.

target recognition is even more restricted — to a 6–8 nt match, the so-called 'seed sequence' of an miRNA[14].

snoRNAs. snoRNPs are involved in the modification of ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs) or tRNAs within the nucleolus of the cell[15,16]. Two families of small ncRNAs — C/D and H/ACA snoRNAs — associate with a core of at least four proteins to form the snoRNP complex: C/D snoRNAs associate with NHPX (also known as NHP2L1, NHP2 non-histone chromosome protein 2-like 1), NOP56 (also known as NOL5A, nucleolar protein 5A), NOP58 and NOP1P (fibrillarin); NOP1P is the modifying enzyme, a methylase that introduces 2′-O-methyl-groups into the riboses of target RNAs (for example, rRNAs, snRNAs or tRNAs)[15]. H/ACA snoRNAs form a protein complex with NHP2P (also known as NOLA2, nucleolar protein family A, member 2), NOP10P (also known as NOLA3), GAR1 (also known as NOLA1) and CBF5P; CBF5P converts uridines into pseudouridines in the same target RNAs as those of C/D box snoRNPs[16]. So the function of the snoRNAs is to guide the enzymes to their respective target sites. This guidance is exerted by short regions (10–20 nt) of complementarity of the guide RNAs to their RNA targets, such as rRNAs, snRNAs or tRNAs (in Archaea only). For example, approximately 200 modification sites are known in mammalian rRNAs, all of them being guided by snoRNAs from the C/D or H/ACA family[15,16].

gRNAs. Another large class of species-specific guide RNAs is present in kinetoplast mitochondria of some parasitic protozoan organisms, such as trypanosomes[17]. These RNAs catalyse the insertion or deletion of U residues into the sequence of many mitochondrial pre-mRNAs, and are designated as gRNAs. gRNAs base pair with mRNA sequences at the U-insertion or deletion site and thereby guide a protein complex, called the editosome, to its proper location. Insertion of U residues involves enzymatic cleavage of the sugar phosphate backbone of the target pre-mRNA, followed by insertion of U residues and subsequent religation of RNA strands, all of which is catalysed by the editosome[18]. The biological function of this RNA-editing mechanism is to generate an ORF within the mitochondrial pre-mRNAs; only after this editing step can these mRNAs be translated into functional proteins[18].

All guide RNA families enable an efficient form of modularity (discussed below in more detail) in which multiple substrates can be processed by a single protein complex. Another type of modularity has recently been observed within the miRNA/siRNA family. Not only can multiple RNAs target a single protein catalytic complex to multiple substrates, but different catalytic enzymes can be transported to different substrates by similar RNP complexes (FIG. 1c). This is observed, for example, in plants, where an miRNA/siRNA–Dicer–Argonaute complex can guide either a DNA methyltransferase[19,20] or histone methyltransferase[21], or an RNA endonuclease[8,10]. What determines the choice of protein partner remains unknown. A similar siRNA/miRNA–RNP complex guides mRNA-translation inhibition in animals[11] and DNA endonucleases in Tetrahymena[22].

Although, according to our definition, an RNP enzyme includes a protein catalyst, we find it useful to extend this concept to include RNA-guided enzymes in which the catalytic part itself is a ribozyme or an RNP. For example, although it is not yet completely clear to what extent the catalytic component of the spliceosome is protein-based[23], it is conceivable that U snRNAs, which target

the splice sites of eukaryotic pre-mRNAs, guide the protein components of the spliceosome complex to their proper locations. Consequently, one might consider the spliceosome to be an example of an RNA-guided enzyme.

A related class of ncRNAs functions through steric hindrance, that is, by competing for a substrate- or an enzyme-binding site, or by changing the structure of a target mRNA. Several members of this class (for example, *DsrA*, *RprA* and *GadY*[24]) are bacterial ncRNAs that are also bound to a common protein component, Hfq. Although these complexes are chimeric RNA–protein enzymes, we do not consider them to be RNA-guided enzymes, because Hfq primarily functions not as a catalyst, but to stabilize the ncRNA[25] or to restructure the ncRNA as an RNA chaperone so it can more effectively interact with its substrate target[26].

Target recognition: RNA versus protein?

DNA or RNA, the natural target molecules of guide RNAs, can also be recognized by protein-only enzymes. In fact, RNA-targeting events that result in base modification of an RNA nucleotide seem to be more often catalysed by protein-only enzymes (A.H., unpublished observations). For example, most tRNA base modifications are catalysed by protein-only enzymes (with few exceptions)[27,28]. Similarly, most examples of RNA editing, in which the coding capacity of an mRNA is altered through base modification, are carried out by protein-only enzymes that belong to either the cytosine deaminase or adenosine deaminase families[29].

A possible explanation for this bias lies in the fact that base pairing, through which guide RNAs recognize their target sites, might interfere with the target's accessibility to the associated protein enzyme. Consistent with this hypothesis, most events that target the RNA backbone, such as cleavage by miRNAs/siRNAs and ribose methylation, are typically carried out by RNA-guided proteins. By contrast, base methylation is generally catalysed by protein-only enzymes. In addition, cleavage of the mature 5′ end of 18S rRNA is guided by U3 snRNA[30], and other cleavage steps in rRNA biogenesis might be guided by U14, U17 and U22 RNAs. Similarly the only form of RNA editing in which guide RNAs are used — the insertion or deletion of U residues within many trypanosome mitochondrial pre-mRNAs — does not involve direct base modification either.

Table 2 | **Principal mechanisms by which ncRNAs function**

Mechanism	Examples	References
Catalysis RNase P → tRNA → Mature tRNA	RNase P, rRNA	44,45,52,53
Guiding siRNP (RISC) + mRNA → Cleaved mRNA	miRNAs, siRNAs, snoRNAs	See TABLE 3
Catalytic cofactor SRP + Ribosome → Stalled ribosome	SRP RNA	1
Antisense recognition *DsrA* RNA + *RpoS* mRNA → Activation of translation	*MicC, DsrA, RprA, GadY, Rev-ErbAα, CopA*	24,54,55,56
Protein binding-site antagonists or agonists 6S RNA + σ⁷⁰ cofactor RNA polmerase → Transcription inhibition of σ⁷⁰ promoters	7SK RNA, 6S RNA, *CsrB/CsrC*	3,4,5
Templating Telomerase + Chromosome DNA ends → DNA synthesis	Telomerase RNA	7

miRNA, microRNA; ncRNA, non-protein-coding RNA; rRNA, ribosomal RNA; siRNA, small interfering RNA; snoRNA, small nucleolar RNA; SRP, signal recognition particle.

This rule is not without exceptions: most pre-rRNA processing events that involve backbone cleavage are probably predominantly carried out by protein-only enzymes, and RNA editing in the mitochondria of myxomycetes (a phylum of fungus-like organisms), which primarily involves base insertion and/or deletion, does not seem to involve any guide RNAs either (at least none has been found so far)[31].

Another exception to the rule that base modifications do not involve guide RNAs is the selective conversion of uridines to pseudouridines in rRNAs and snRNAs, which in Eukarya and Archaea is typically guided by H/ACA snoRNAs[16]. Interestingly, however, in this case the snoRNA target sites are restricted to sequences that lie 5′ and 3′ to the modified base and do not include the targeted base

itself, thereby forming a pseudouridylation pocket (FIG. 2). This spatial arrangement therefore might not interfere with the action of the modifying enzyme, the CBF5P pseudouridylase.

Pros and cons of RNA guiding

Given that RNA or DNA target recognition can also be accomplished by proteins alone, why is the RNA-guided enzyme mechanism so widely used?

A possible explanation that has been implicated previously[32] and that has become more attractive as more guide RNAs and guide-RNA families are identified is based on the observation that an RNA-guided enzyme system requires only one (non-sequence-specific) protein for its enzymatic activity. Sequence specificity, and thereby target recognition,

Table 3 | **Guide RNAs, their catalytic payloads and function**

ncRNAs	Number of ncRNAs	Catalytic payload	Cellular function	References
Experimentally established guide RNAs with protein catalysts				
miRNAs	>400*	RNA induced silencing complex	Translational control; mRNA destabilization	11–13,57,58
siRNAs	N.D.	DNA methyltransferase	Transcriptional control; transposon or viral protection	20,59
rasiRNAs	N.D.	Unknown: transcriptional or post-transcriptional regulation	Repeat silencing	60
Heterochromatic siRNAs	N.D.	Histone methyltransferase	Heterochromatin silencing	21,61,62
siRNAs	N.D.	RNA endonuclease	mRNA degradation	8
H/ACA snoRNAs	~100*	RNA pseudouridylase	rRNA and snRNA modification	16,63
C/D snoRNAs	~100*	RNA methyltransferase	rRNA, snRNA and tRNA (in Archaea only) modification	15,63
U3 (C/D snoRNA)	1*	RNA endonuclease	5′ pre-rRNA processing	30
U7 (snRNA)	1	RNA endocuclease	3′ end processing of histone pre-mRNAs	64,65
Editing gRNAs	50–100‡	U insertion or deletion enzymes	Render mRNAs fully translatable in trypanosome mitochondria	18
Putative guide RNAs				
Xist	1*	DNA methyltransferase	Dosage compensation	66,67
U snRNAs	5*	Spliceosome RNP	Splicing	68,69

*Number estimated in humans. ‡Number estimated in trypanosomes. gRNA, classical guide RNA; miRNA, microRNA; ncRNA, non-protein-coding RNA; N.D., not determined; rasiRNA, repeat-associated siRNA; rRNA, ribosomal RNA; siRNA, small interfering RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA; Xist, X-(inactive)-specific transcript.

is accomplished by the small ncRNA component of the RNP complex. This strategy both limits the amount of the genome that needs to be allocated to encode the required genes and facilitates the evolution of novel targets for the complex.

For example, if only protein enzymes were used to catalyse the approximately 200 observed mammalian rRNA methylation and pseudouridylation events, as many as 200 proteins would have to be synthesized. Similarly, because as many as one-third of mammalian mRNAs are estimated to be targets of post-transcriptional gene regulation by miRNAs[33], many hundreds or even thousands of individual proteins would be required if each regulatory protein had only a single target. Protein-only enzymes also show some flexibility in target recognition. For example, tRNA-modification enzymes can recognize and modify conserved sequence or structure motifs within different tRNAs[27,28]. Nevertheless, this target-selection flexibility is far more limited and constrained than that of RNA-guided enzymes.

Moreover, evolutionary mechanisms that generate new targets for protein-only enzymes are necessarily more complex. This is because new genes rarely arise *de novo*, but rather by gene duplication, followed by mutation of the duplicated copy[34]. To accomplish recognition of new target sites, a sophisticated mutation mechanism would be required. This mechanism would probably require multiple point mutations, changing several amino acids, in order to modify the RNA-binding domain to target the new site. Because many protein mutations within RNA-binding domains would be expected to result in loss of function, this would be a highly inefficient means of generating target diversity.

By contrast, RNA-guided systems avoid both the problem of requiring multiple protein enzymes to catalyse reactions that involve multiple substrates and the difficulties of evolving other enzymes for new target sites. A single RNA-guided protein catalytic complex can carry out many modifications or cleavages simply by associating with the appropriate guide RNA. Because guide RNA genes are generally much shorter than protein-coding genes, significant gains in genomic coding efficiency are possible. In addition, the energy cost of synthesizing a protein molecule is much higher than for an RNA molecule.

Furthermore, the RNA-guided enzyme system has the potential to expand its repertoire of target sites simply by duplicating the gene for the RNA guide and incorporating single nucleotide mutations within its antisense sequence. Such single-base antisense mutations will more often generate a new set of target sites and will rarely lead to loss of functionality of the RNP complex, compared with mutations in protein genes. Evidence of such generation of a novel guide by gene duplication has recently been detected in the *Arabidopsis thaliana* genome[35].

Using RNA as the basis for a nucleic-acid target recognition makes it relatively easy to modulate specificity during evolution. For example, in a genome with 25,000 transcripts of an average length of 4 kb there are $\sim100 \times 10^6$ potential post-transcriptional target binding sites. Assuming uniform transcript-base composition, a guide RNA of length N would exactly base pair at $10^8/4^N$ locations in the transcriptome. Consequently, a 13 nt guide RNA would typically have a single target site, and changing a single nucleotide in the 13 nt would result in the targeting of a different unique location. If less specificity and more diversity are required, 10 nt RNA guides could be used. In this case each guide would on average have ~100 target sites and a single nucleotide change would result in 100 different targets. To achieve this flexibility of target selection with protein-only enzymes would probably require a much larger number of nucleotide substitutions, with an increased likelihood that at least one of these mutations would cause a loss of function.

Given the above, it might be expected that species without an ncRNA-guided system for specific RNA or DNA targeting will contain fewer modification target
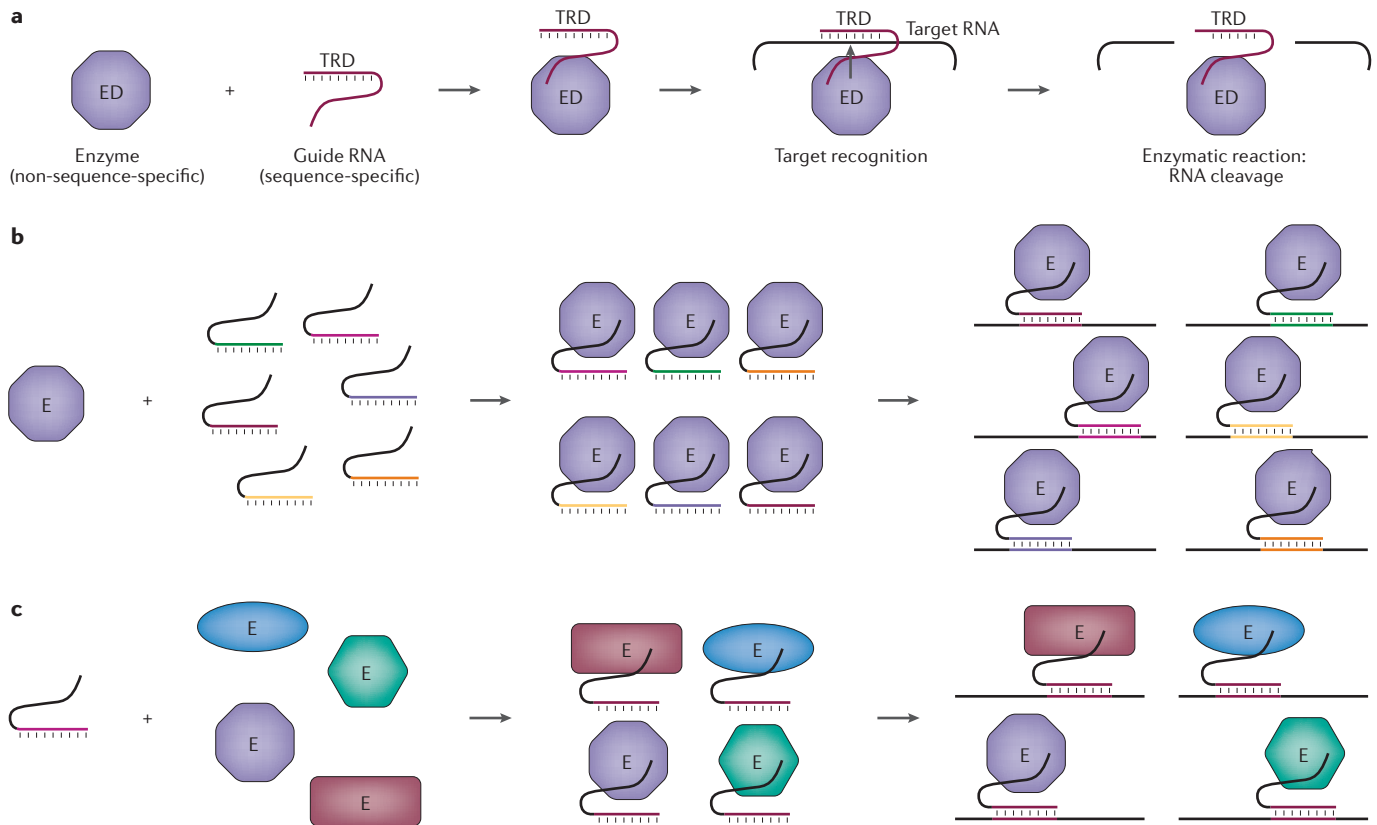
Figure 1 | **The concept of RNA guiding. a** | A non-sequence-specific enzyme complex (which consists of one or more proteins) associates with a guide RNA that provides the target specificity; this chimeric RNA–protein complex is then able to target RNA or DNA molecules to exert its enzymatic function (for example, RNA cleavage, as in the case of a small interfering RNA–protein complex, the RNA induced silencing complex). ED, enzyme domain; TRD, target recognition domain. **b** | One enzyme complex (E), which consists of one to several proteins, binds to many small guide RNAs that recognize their targets by Watson–Crick base pairing and thereby guide the enzymatic complex to different substrates. **c** | One guide RNA can recognize different enzyme complexes (E): one class of classical guide RNAs (such as microRNAs) might bind to different enzyme complexes (E) and so is able to guide different enzymatic reactions.

sites, such as cleavage sites or methylation/pseudouridylation sites. This is observed for the modification of rRNAs. Bacteria, which lack the snoRNA-based modification system, have few rRNA modifications[36]. By contrast, Eukarya, which have snoRNA-guided systems, have several hundred rRNA-modification sites[15,16].

Considering the efficiency and target diversity of RNA-guided systems, why have they not been adopted more widely? For example, why have snoRNA-guided RNA-modification systems not evolved in bacteria? One possible reason is that the increase in the number of potential target sites might be disadvantageous, for example, if the modifications were applied promiscuously without proper regulation. It has recently been shown that snoRNA-guided rRNA ribose methylation can lead to growth defects in yeast if inappropriate nucleotides are methylated[37]. Therefore, expression of RNA-guided enzymes might not have been adopted more widely owing

to the deleterious effects that they could have. The localization of RNA-guided modifications in eukaryotic cells[16] to specific sub-cellular components might have evolved to protect the cell from such potentially deleterious effects of promiscuous snoRNA-guided RNA modification. Bacteria that lack such cellular substructures might be more susceptible to undesirable mRNA modifications that could be mediated by the guide RNA-based system itself and so have not developed RNA-guided enzyme machinery.

Interestingly, Archaea, which lack a nucleus-like structure but possess an RNA-guided rRNA-modification system, seem to be an intermediate between Bacteria and Eukarya in terms of numbers of rRNA modifications and guide RNA species. Non-thermophilic Archaea have relatively few C/D or H/ACA snRNAs, whereas thermophilic ones also have relatively few H/ACA snRNAs but higher numbers of C/D snRNAs[38,39].

RNA-guiding trade-offs
These trade-offs indicate that some classes of organisms might prefer RNA-guided mechanisms whereas for others selection would favour protein-only mechanisms. However, in some cases the reasons for the selection of one mode of substrate targeting over the other is less clear. Sometimes the same function might be carried out by a protein-only enzyme in one species, but by an RNA-guided protein enzyme in a close relative. For example, in all but one archaeal species investigated so far, the 2′-O-methylation at position C_{56} within certain tRNAs is carried by a protein-only enzyme (as is also the case in eukaryotes). However, in the archaeal species *Pyrobaculum aerophilum* this tRNA-modifying enzyme is missing and instead the modification is carried out by a C/D snoRNP[40].

Two other intriguing examples of the coexistence of multiple RNA-modification mechanisms come from *Saccharomyces cerevisiae*. Yeast U2 snRNA is
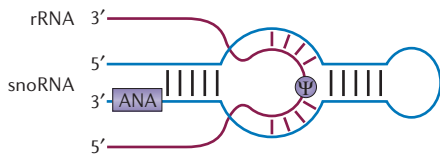
Figure 2 | **Pairing of a H/ACA snoRNA with an rRNA site for pseudouridine formation.** Targeting involves base pairing of the guide RNA with complementary ribosomal RNA (rRNA) sequences that flank the uridine to be modified (Ψ). The RNA duplex that is formed between the small nucleolar RNA (snoRNA) and the rRNA excludes the modified base itself, thereby rendering it accessible to the modifying enzyme CBF5P. Reproduced with permission from REF. 70 © (1998) American Society for Microbiology.

pseudouridylated at positions 35 (Ψ35), 42 (Ψ42) and 44 (Ψ44). It has recently been demonstrated that although Ψ35 and Ψ44 modifications are catalysed by a protein-only enzyme, modification of Ψ42 is catalysed by a snoRNP[41]. In the second example, 2′-O-methylation of yeast rRNA is generally mediated by a C/D box snoRNA-guided system but 2′-O-methylation of G2922 within 28S rRNA is mediated by a protein-only enzyme, Spb1p (REF. 42). So, even within the same organism, a single RNA substrate can be modified at different target sites by two distinct mechanisms: one using a protein-only enzyme and the other an RNA-guided enzyme.

How did RNA-guided enzymes evolve?

We can imagine three alternative models of how primordial ribozymes were replaced by RNA–protein enzymes during evolution (FIG. 3).

The target recognition domain (TRD) of the RNA enzyme could be replaced by a protein; alternatively, the enzyme domain (ED) of the original ribozyme could be replaced by, presumably, a more potent protein domain, whereas target recognition could still rely on the RNA portion. Subsequent replacement of the RNA TRD or RNA ED would result in protein-only enzymes. Finally, but less likely, the replacement of the ribozyme by a protein-only enzyme could involve a one-step mechanism (FIG. 3).

In the stepwise evolutionary replacement of EDs or TRDs, it seems more likely that the enzyme domain would be replaced by a protein than the target domain. This is because proteins, which are made up of 21 different amino acids (including selenocysteine), are likely to be catalytically

more efficient than RNAs, which consist of only 4 different nucleotides. The rationale behind this assumption is that the 21 side chains of amino acids might be able to catalyse more diverse chemical reactions than the 4 bases. This expectation is consistent with what is observed: although there are many RNA-guided proteins, very few truly transacting ribozymes, such as RNase P RNA (which catalyses cleavage of precursor tRNAs[43]) or large rRNA (which catalyses peptide-bond formation[44–46]) have been identified so far.

It is also possible — and indeed likely — that different RNA-guided enzymes developed at various time-points during evolution. For example, because gRNAs that mediate U insertion or deletion are found in such a restricted class of species it seems likely that they are of relatively recent evolutionary origin. By contrast, the snoRNAs, for example, probably evolved before the divergence of Eukarya and Archaea[47], but after the divergence from Bacteria, because this domain of life has so far not been shown to contain snoRNAs.

In general, it seems that guide RNAs are mainly — if not exclusively — found in

Eukarya and Archaea, and therefore appear to be a rather recent evolutionary acquisition (TABLE 3). By contrast, in Bacteria, which also use antisense RNAs (in analogy to guide RNAs), no enzyme is guided to the target. Instead, the antisense RNA itself appears to exert the function in the absence of the enzyme (for example, *DsrA* and *OxyS* RNAs, see above). Archaea seem to represent a middle ground between Eukarya and Bacteria — they contain *bona fide* guide RNAs, such as snoRNAs, but also contain a considerable number of bacterial antisense-like RNAs[50,51].

It should be noted that protein or protein–RNA enzymes might have already existed in the earliest life forms. It is therefore conceivable that small proteins or peptides evolved in parallel with RNA macromolecules, before the evolution of ribozymes, and that, at some point in evolution, interactions between the two macromolecules might have been synergistic, ultimately resulting in the evolution of RNA–protein enzymes (RNPs). The problem with this model is that the puzzle remains about how these primordial peptides and proteins evolved and were passed from generation to generation. Presumably
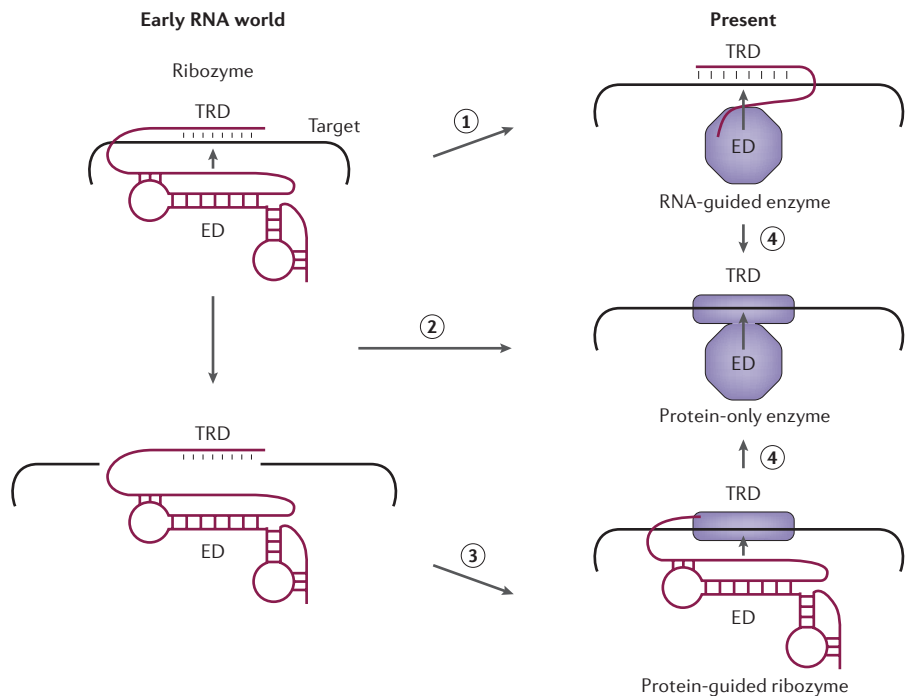

Figure 3 | **A putative model of RNA-guided and protein-only enzyme evolution.** Three routes of evolution can be predicted from an early RNA-world ribozyme that is able to target and cleave an RNA molecule. **1** | Replacement of the enzyme domain (ED) of the ribozyme by a protein, while leaving the target recognition domain (TRD) as an RNA; examples include micro and small interfering ribonucleo-proteins (miRNPs and siRNPs). **2** | Conversion from the ribozyme to the protein enzyme in a single step; examples include RNase Z and RNase A. **3** | Replacement of the TRD by a protein domain, while leaving the ED as an RNA; examples include RNase P. **4** | Subsequent replacement of either the RNA TRD or RNA ED in these RNA–protein enzymes results in a protein-only enzyme.

some form of protein self-replication would be required to achieve this. There is currently no evidence to indicate how this might have occurred.

Conclusion

The first RNAs to be called guide RNAs were those found in kinetoplast mitochondria of trypanosomes, which guide the insertion or deletion of U residues into mitochondrial pre-mRNAs[18]. Here we suggest that the concept of guide RNAs is far more widespread than initially anticipated and can be extended to snoRNAs, siRNAs/miRNAs and even snRNAs. (Indeed, two other ncRNA families have recently been identified in *Caenorhabditis elegans*[48]. Whether these ncRNAs represent new guide-RNA families is unknown although it seems likely that our list of guide-RNA families is still incomplete.) The large families of guide RNAs outnumber the relatively few representatives of catalytic ncRNAs, to which most attention has been drawn in recent years. In evolutionary terms, especially in eukaryotes, the concept of RNA guiding has proved a powerful way of generating genetic diversity because new target sites can be generated by gene duplication of guide-RNA genes and mutation of their antisense elements.

How will guide RNAs and their associated enzymes evolve in the future? One model predicts that in the distant future the RNA target recognition domain of the chimeric RNA–protein enzyme will be replaced by a protein, able to guide it to the respective target site (FIG. 3). But evolution might not work that way because ncRNAs are potent in selectively recognizing other nucleic acids by sequence-specific base pairing. Therefore, the number of RNA-guided enzymes might in fact expand in the future, leading to an even more elaborate regulatory system in eukaryotic cells. Furthermore, many RNA-guided systems might be caught in an evolutionary trap. For example, the RNA-guided mechanism might already be so far evolved that it would be impossible to replace it by a protein-only based mechanism, without losing specificity and/or biological function.

Alexander Hüttenhofer is at the Innsbruck Biocenter, Medical University Innsbruck, Fritz-Pregl-Strasse 3, 6020 Innsbruck, Austria.

Peter Schattner is at the Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95060, USA.

Correspondence to A.H.
e-mail: Alexander.Huettenhofer@i-med.ac.at

1. Halic, M. & Beckmann, R. The signal recognition particle and its interactions during protein targeting. *Curr. Opin. Struct. Biol.* **15**, 116–125 (2005).
2. Halic, M. *et al.* Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* **427**, 808–814 (2004).
3. Yang, Z., Zhu, Q., Luo, K. & Zhou, Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414**, 317–322 (2001).
4. Wassarman, K. M. & Storz, G. 6S RNA regulates *E. coli* RNA polymerase activity. *Cell* **101**, 613–623 (2000).
5. Dubey, A. K., Baker, C. S., Romeo, T. & Babitzke, P. RNA sequence and secondary structure participate in high-affinity CsrA–RNA interaction. *RNA* **11**, 1579–1587 (2005).
6. Pauler, F. M., Stricker, S. H., Warczok, K. E. & Barlow, D. P. Long-range DNase I hypersensitivity mapping reveals the imprinted *Igf2r* and *Air* promoters share *cis*-regulatory elements. *Genome Res.* **15**, 1379–1387 (2005).
7. Blackburn, E. H. Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. *FEBS Lett.* **579**, 859–862 (2005).
8. Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343–349 (2004).
9. Filipowicz, W. RNAi: the nuts and bolts of the RISC machine. *Cell* **122**, 17–20 (2005).
10. Tomari, Y. & Zamore, P. D. Perspective: machines for RNAi. *Genes Dev.* **19**, 517–529 (2005).
11. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437–1441 (2004).
12. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
13. Pillai, R. S., Artus, C. G. & Filipowicz, W. Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis. *RNA* **10**, 1518–1525 (2004).
14. Farh, K. K. *et al.* The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310**, 1817–1821 (2005).
15. Bachellerie, J. P., Cavaille, J. & Huttenhofer, A. The expanding snoRNA world. *Biochimie* **84**, 775–790 (2002).
16. Kiss, T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* **109**, 145–148 (2002).
17. Blum, B., Bakalara, N. & Simpson, L. A model for RNA editing in kinetoplastid mitochondria: 'guide' RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* **60**, 189–198 (1990).
18. Stuart, K. D., Schnaufer, A., Ernst, N. L. & Panigrahi, A. K. Complex management: RNA editing in trypanosomes. *Trends Biochem. Sci.* **30**, 97–105 (2005).
19. Kawasaki, H. & Taira, K. Induction of DNA methylation and gene silencing by short interfering RNAs in human cells. *Nature* **431**, 211–217 (2004).
20. Matzke, M. A. & Birchler, J. A. RNAi-mediated pathways in the nucleus. *Nature Rev. Genet.* **6**, 24–35 (2005).
21. Volpe, T. A. *et al.* Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**, 1833–1837 (2002).
22. Mochizuki, K. & Gorovsky, M. A. Small RNAs in genome rearrangement in Tetrahymena. *Curr. Opin. Genet. Dev.* **14**, 181–187 (2004).
23. Butcher, S. E. & Brow, D. A. Towards understanding the catalytic core structure of the spliceosome. *Biochem. Soc. Trans.* **33**, 447–449 (2005).
24. Gottesman, S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.* **21**, 399–404 (2005).
25. Zhang, A. *et al.* Global analysis of small RNA and mRNA targets of Hfq. *Mol. Microbiol.* **50**, 1111–1124 (2003).
26. Geissmann, T. A. & Touati, D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J.* **23**, 396–405 (2004).
27. Hopper, A. K. & Phizicky, E. M. tRNA transfers to the limelight. *Genes Dev.* **17**, 162–180 (2003).
28. Agris, P. F. Decoding the genome: a modified view. *Nucleic Acids Res.* **32**, 223–238 (2004).
29. Lehmann, K. A. & Bass, B. L. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. *Biochemistry* **39**, 12875–12884 (2000).
30. Dragon, F. *et al.* A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* **417**, 967–970 (2002).
31. Horton, T. L. & Landweber, L. F. Rewriting the information in DNA: RNA editing in kinetoplastids and myxomycetes. *Curr. Opin. Microbiol.* **5**, 620–626 (2002).
32. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929 (2001).
33. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nature Genet.* **37**, 495–500 (2005).
34. Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature Rev. Genet.* **4**, 865–875 (2003).
35. Allen, T. A., Von Kaenel, S., Goodrich, J. A. & Kugel, J. F. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. *Nature Struct. Mol. Biol.* **11**, 816–821 (2004).
36. Ofengand, J. Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett.* **514**, 17–25 (2002).
37. Liu, B. & Fournier, M. J. Interference probing of rRNA with snoRNPs: a novel approach for functional mapping of RNA *in vivo*. *RNA* **10**, 1130–1141 (2004).
38. Tang, T. H. *et al.* Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl Acad. Sci. USA* **99**, 7536–7541 (2002).
39. Dennis, P. P., Omer, A. & Lowe, T. A guided tour: small RNA function in Archaea. *Mol. Microbiol.* **40**, 509–519 (2001).
40. Renalier, M. H., Joseph, N., Gaspin, C., Thebault, P. & Mougin, A. The Cm56 tRNA modification in Archaea is catalyzed either by a specific 2′-O-methylase, or a C/D sRNP. *RNA* **11**, 1051–1063 (2005).
41. Ma, X. *et al.* Pseudouridylation of yeast U2 snRNA is catalyzed by either an RNA-guided or RNA-independent mechanism. *EMBO J.* **24**, 2403–2413 (2005).
42. Lapeyre, B. & Purushothaman, S. K. Spb1p-directed formation of Gm2922 in the ribosome catalytic center occurs at a late processing stage. *Mol. Cell* **16**, 663–669 (2004).
43. Kirsebom, L. A. RNase P — a 'Scarlet Pimpernel'. *Mol. Microbiol.* **17**, 411–420 (1995).
44. Noller, H. F., Hoffarth, V. & Zimniak, L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **256**, 1416–1419 (1992).
45. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930 (2000).
46. Polacek, N. & Mankin, A. S. The ribosomal peptidyl transferase center: structure, function, evolution, inhibition. *Crit. Rev. Biochem. Mol. Biol.* **40**, 285–311 (2005).
47. Tran, E., Brown, J. & Maxwell, E. S. Evolutionary origins of the RNA-guided nucleotide-modification complexes: from the primitive translation apparatus? *Trends Biochem. Sci.* **29**, 343–350 (2004).
48. Deng, W. *et al.* Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res.* **16**, 20–29 (2006).
49. Huttenhofer, A., Brosius, J. & Bachellerie, J. P. RNomics: identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.* **6**, 835–843 (2002).
50. Tang, T. H. *et al.* Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **55**, 469–481 (2005).
51. Zago, M. A., Dennis, P. P. & Omer, A. D. The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol Microbiol.* **55**, 1812–1828 (2005).
52. Doudna, J. A. & Cech, T. R. The chemical repertoire of natural ribozymes. *Nature* **418**, 222–228 (2002).
53. Doudna, J. A. & Rath, V. L. Structure and function of the eukaryotic ribosome: the next frontier. *Cell* **109**, 153–156 (2002).
54. Storz, G., Altuvia, S. & Wassarman, K. M. An abundance of RNA regulators. *Annu. Rev. Biochem.* **74**, 199–217 (2005).
55. Hastings, M. L., Ingle, H. A., Lazar, M. A. & Munroe, S. H. Post-transcriptional regulation of thyroid hormone receptor expression by *cis*-acting sequences and a naturally occurring antisense RNA. *J. Biol. Chem.* **275**, 11507–11513 (2000).
56. Nordstrom, K. Plasmid R1 — replication and its control. *Plasmid* **55**, 1–26 (2006).
57. Ambros, V. MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* **113**, 673–676 (2003).

58. Lim, L. P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
59. Chan, S. W. *et al.* RNA silencing genes control *de novo* DNA methylation. *Science* **303**, 1336 (2004).
60. Aravin, A. A. *et al.* The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* **5**, 337–350 (2003).
61. Reinhart, B. J. & Bartel, D. P. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**, 1831 (2002).
62. Cam, H. P. *et al.* Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome. *Nature Genet.* **37**, 809–819 (2005).
63. Decatur, W. A. & Fournier, M. J. RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.* **278**, 695–698 (2003).
64. Dominski, Z. & Marzluff, W. F. Formation of the 3′ end of histone mRNA. *Gene* **239**, 1–14 (1999).
65. Marzluff, W. F. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr. Opin. Cell Biol.* **17**, 274–280 (2005).
66. Nusinow, D. A. & Panning, B. Recognition and modification of sex chromosomes. *Curr. Opin. Genet. Dev.* **15**, 206–213 (2005).
67. Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the mammalian X chromosome. *Annu. Rev. Genomics Hum. Genet.* **6**, 69–92 (2005).
68. Guthrie, C. & Patterson, B. Spliceosomal snRNAs. *Annu. Rev. Genet.* **22**, 387–419 (1998).
69. Sharp, P. A. The discovery of split genes and RNA splicing. *Trends Biochem. Sci.* **30**, 279–281 (2005).
70. Ofengand, J. & Fournier, M. J. in *Modification and Editing of RNA* (eds Grosjean, H. & Benne, R.) Ch. 12 (American Society for Microbiology Press, Washington DC, 1998).

OPINION

Addressing the problems with life-science databases for traditional uses and systems biology

Stephan Philippi and Jacob Köhler

Abstract | A prerequisite to systems biology is the integration of heterogeneous experimental data, which are stored in numerous life-science databases. However, a wide range of obstacles that relate to access, handling and integration impede the efficient use of the contents of these databases. Addressing these issues will not only be essential for progress in systems biology, it will also be crucial for sustaining the more traditional uses of life-science databases.

Several decades ago, scientists started to set up biological data collections for the centralized management of and easy access to experimental results, and to ensure long-term data availability (FIG. 1a). Many early data collections were initially administered using word processing or spreadsheet applications. Owing to the limited amount of data that could be stored in this way, and the reductionist viewpoint that characterized most biological research at that time, this approach to data collection seemed reasonable, and was sufficient for occasional exchanges with colleagues.

However, with the exponential growth of experimental data that is taking place owing to rapid biotechnological advances and high-throughput technologies, as well as the advent of the World Wide Web as a new means for data exchange, the world dramatically changed. The huge amounts of data that are now produced on a daily basis require more sophisticated management solutions, and the availability of the internet as a modern infrastructure for scientific exchange has created new demands with

respect to data accessibility. Furthermore, the relatively new field of systems biology has further increased the requirements that are demanded of life-science databases. The general vision of systems biology is to move out of the era of reductionist studies of isolated parts of interest — for example, individual proteins and genes — and to develop a molecular understanding of more complex structures and their dynamics, such as regulatory networks, cells, organs and, ultimately, whole organisms[1].

The most important tool for reaching an understanding of biology at the level of systems is the analysis of biological models (FIG. 1b). The basic building blocks for these models are existing experimental data, which are stored in literally thousands of databases[2–4]. As a result, database integration is a fundamental prerequisite for any study in systems biology[5,6]. Because database integration has long been recognized as a key technology in the life sciences, research in this area also has a long tradition. However, although many approaches exist, database integration in the life sciences is still far from being trivial.

A common misconception is that the main problems of database integration are related to the technology that is used for these purposes. Here we argue that although the mastering of such technology can be challenging, the main problems are actually related to the databases themselves. There are many issues with life-science databases that prevent the effective use of integration technology. These problems not only have adverse effects on the quintessential task of ensuring data availability to the general research community, but present an even greater obstacle to systems biology. Here we provide a systematic analysis of the common problems that relate to life-science databases — which are technical, social and political — and suggest solutions for how they could be overcome.

Technical problems
As a prerequisite for the discussion of technical problems with life-science databases it is important to understand the general principles of database integration. Life-science databases have experienced an exponential growth in numbers in recent years and contain information of many types[7]. To bridge the gap between these often unconnected islands of biological knowledge, and between the different types of experimental data that they contain, various approaches to data integration have been pursued over the past decade. These range from basic hypertext linking to more advanced approaches that involve the use of federated databases and data warehouses (BOX 1). It is on the advanced approaches that we focus here, as they provide the best illustration of the diverse problems with life-science databases that affect data integration, particularly with respect to the goals of systems biology.

Although there are many variants of the more advanced applications, the problems with life-science databases that affect integration using federated database technology or data warehouses are almost identical.