size difference and local mutation rate do show covariance[6].

An alternative possibility is that recombination, proportional GC content and intron size do show covariance, but not in the same way in all vertebrates. It might be that in chickens and mammals there is more recombination in GC-rich regions, but in the cold-blooded species the opposite pattern is found. We are unaware of pertinent data to test the prediction with respect to *Xenopus* and *Fugu* (or indeed any other cold-blooded vertebrate). However, we can provide a test of the hypothesis in the case of chickens. If proportional GC content and recombination rate were related in the above manner, then the Z chromosome should have a GC content that is lower than that of autosomal genes, because the former recombines in males alone (except in the pseudoautosomal region). The W chromosome should have an even lower figure because it never recombines.

Chickmap (www.ri.bbsrc.ac.uk/chickmap/) provides information on the map position of chicken genes. From here we have derived a list of genes whose full cDNA was known and that were also known to be either Z-linked (N = 7) or autosomal (N = 23). We find that the Z-linked genes have a mean proportional GC3 content of 41% compared with 67% for autosomal sequences. These figures are highly significantly different (in the Mann-Whitney U test, $P = 0.0006$). Our figures appear to be consistent with analysis of all 1454 chicken coding sequences described in GenBank. The mean proportional GC3 content for these is 60.4% (from www.dna.affrc.go.jp/~nakamura/CUTG.html). It is to be expected that our autosomal figure should be above this because some (unknown) proportion of the 1454 genes are Z-linked.

Only four sequences with putative open reading frames have been described on the chicken W-chromosome. With a mean proportional GC3 content of 34.7%, these have, as expected, a mean GC3 content lower than both Z-linked and autosomal genes. However, whereas this figure is significantly lower than the autosomal figure (in the Mann-Whitney U test, $P = 0.0009$), it is not significantly different from the sequences on the Z-chromosome (in the Mann-Whitney U test, $P = 0.149$), although with a total sample size of only 11, this should not be taken as a strong rejection.

We conclude that, at least in some warm-blooded vertebrate species, there is a significant tendency for introns to be smaller in GC-rich regions. We have failed to reject the hypothesis that, in these species, the recombination rate also positively covaries with GC content. Hence, we cannot reject the hypothesis that recombination explains some (but possibly not much) of the variation in intron size, but we cannot know whether this is the result of stronger selection that is associated with recombination, or associated with a mutational bias. Analysis of the recombination pattern in cold-blooded species will provide a further test of the proposed link between recombination and intron size.

Acknowledgements

We thank two anonymous referees for their comments and L. Duret for access to unpublished data.

References

1 Duret, L. *et al.* (1995) Statistical-analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* 40, 308–317
2 Eyre-Walker, A. (1993) Recombination and mammalian genome evolution. *Proc. R. Soc. London B* 252, 237–243
3 Nordborg, M. *et al.* (1996) The effect of recombination on background selection. *Genet. Res.* 67, 159–174
4 Duret, L. *et al.* (1994) Hovergen – a database of homologous vertebrate genes. *Nucleic Acids Res.* 22, 2360–2365
5 Bernardi, G. *et al.* (1997) The major compositional transitions in the vertebrate genome. *J. Mol. Evol.* 44, 44–51
6 Ogata, H. *et al.* (1996) The size differences among mammalian introns are due to the accumulation of small deletions. *FEBS Lett.* 390, 99–103

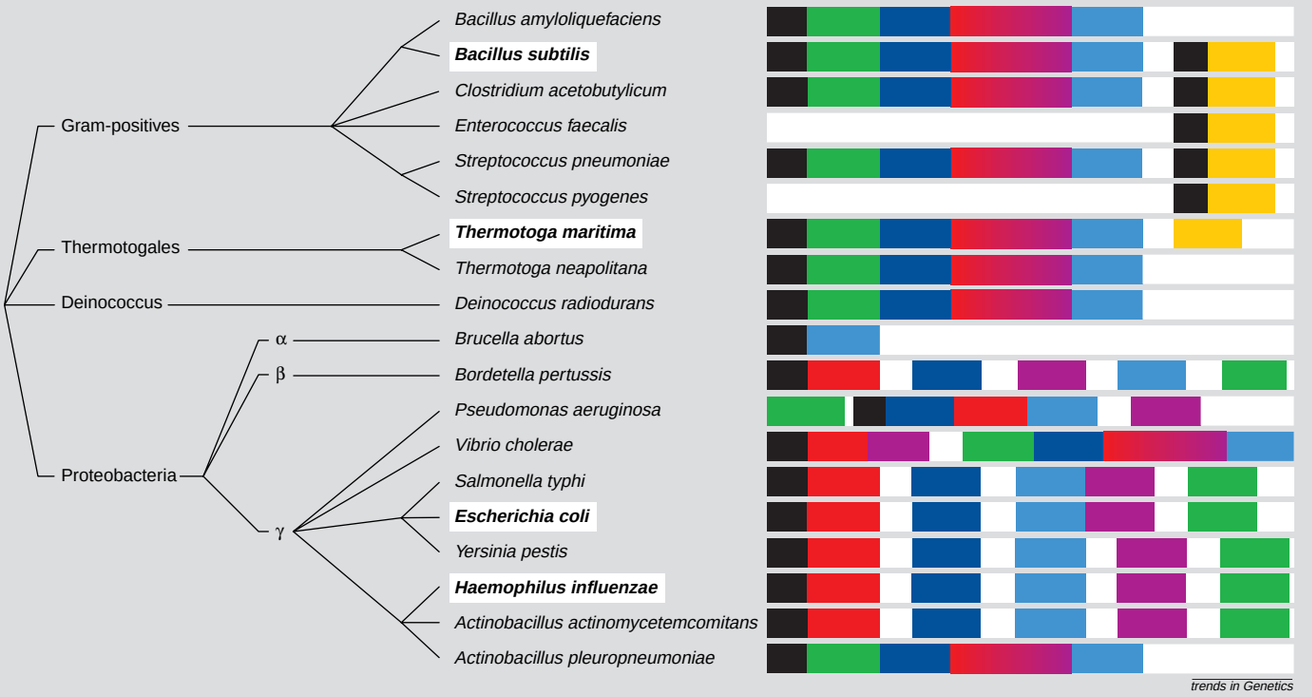
A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes

Large-scale sequencing of bacterial genomes has opened a new era in computational genomics. Gene complements are successfully analysed on the protein level. However, it has been noted that regulatory sites are less conserved than genes[1], although it is possible to use genomic comparisons in order to predict gene regulation at the level of DNA (Ref. 2) and RNA (Refs 3, 4). Indeed, comparative analysis is one of the standard methods to predict the secondary structure of RNA. It has been used to determine the spatial structure of stable RNAs (Refs 5–7), and to analyse regulatory RNAs, mostly in viral genomes[8,9]. In such studies, either the common fold is selected among many predicted suboptimal folds for a set of RNAs (Ref. 8), or analysis of complementary substitutions in aligned sequences is used to construct a single conserved structure[7,9]. Non-viral regulatory RNA elements, such as iron-responsive elements[10], and sites that regulate the initiation of translation in operons of ribosomal proteins[3,4,11] often involve conserved structural elements and conserved nucleotides.

Previously, we have identified a regulatory region upstream of the riboflavin operon of *Bacillus subtilis* and *Bacillus amyloliquefaciens*. Mutations in this region influence the level of riboflavin synthesis[12,13]. Surprisingly, this

0168-9525/99/$ – see front matter PII: S0168-9525(99)01856-9

FIGURE 1. Phylogeny of the genomes harboring the *RFN* element and riboflavin biosynthesis operons



Text in white boxes indicates complete genomes. Traditional gene names are different in Gram-positive and Gram-negative bacteria, so the genes are shown by colored bars: green, riboflavin-specific deaminase (*ribG* / *ribD*); dark blue, riboflavin synthase alpha subunit (*ribB* / *ribE*); light blue, riboflavin synthase beta subunit (*ribH* / *ribH*); red, 3,4-dihydroxy-2-butanone 4-phosphate synthase (*ribF* / *ribB*); magenta, GTP cyclohydrolase (*ribA* / *ribA*); yellow, *ypaA*; black, *RFN* element.

region is strongly conserved in diverse branches of bacteria (Fig. 1). It can fold into a conserved structure with five hairpins (*RFN* element), which has some features that make it unique among other conserved RNA elements.

In Gram-positive bacteria, *Thermotoga* spp., *Deinococcus radiodurans*, and in one proteobacterium, *Actinobacillus pleuropneumoniae*, the riboflavin synthesis genes form a single operon with the same order of genes (riboflavin-specific deaminase and reductase *ribG*; riboflavin synthase alpha subunit *ribB*; fused genes encoding 3,4-dihydroxy-2-butanone 4-phosphate synthase and GTP cyclohydrolase II *ribFA*, and riboflavin synthase beta subunit *ribH*). The *RFN* element is upstream of the riboflavin operon and in *Bacillus* spp. and *A. pleuropneumoniae*, where the transcription start is known, the *RFN* element lies within the transcribed region[12–14]. Genomes of the Gram-positive bacteria contain the second copy of the *RFN* element upstream of gene *ypaA* (in *B. subtilis* nomenclature) encoding an unknown protein with five candidate transmembrane segments. The genome of *Thermotoga maritima* contains a distant ortholog of *ypaA* (*TM1455*) that is not associated with the *RFN* element.

Traditional gene names are different in Gram-negative bacteria. In *Pseudomonas aeruginosa*, the *RFN* element lies between *ribD* (riboflavin-specific deaminase) and *ribE* (riboflavin synthase alpha subunit). In some ribosomal protein operons, the regulatory element situated between genes still influences the rate of translation of upstream genes[11]. In *Brucella abortus*, the *RFN* element is located upstream of the open reading frame, originally annotated as 'antigen'[15], that is highly similar to the riboflavin synthase beta subunit. In other proteobacteria, the *RFN* element is upstream of *ribB* (3,4-dihydroxy-2-butanone 4-phosphate synthase).

The aligned sequences of the *RFN* elements are given in Fig. 2. The conserved secondary structure has five stem-loops and a single root stem of at least five base pairs (Fig. 3). Out of 20 base pairs, ten are invariant on the sequence level, whereas the remaining positions are confirmed by compensatory substitutions. Out of 47 single-strand positions, 24 are absolutely invariant. The preliminary structure that has been previously published[13] is not supported by genome comparisons.

There are two non-conserved structure elements: an additional stem-loop on the top of stem-loop 2, and a variable stem-loop 3. The former does not seem to be correlated with phylogeny, whereas the exact form of the latter depends on the phylogenetic position of a genome. Indeed, the length of the variable stem-loop 3 is 10–20 nucleotides in Gram-positive bacteria and *Thermotoga* spp. and 25–76 nucleotides in proteobacteria (excluding *A. pleuropneumoniae*, see below) and *D. radiodurans*. In proteobacteria, this stem-loop always starts with unpaired GAGCG and ends with unpaired GTCAGCAGA with a long paired region in between.

The genome of *A. pleuropneumoniae* is different from the genomes of other proteobacteria, because the riboflavin genes form an operon exactly like the one in Gram-positive bacteria. The variable stem-loop 3 of the *RFN* element is shorter than in other proteobacteria, and it does not have unpaired GAGCG and GTCAGCAGA as all other proteobacterial *RFN* elements do. Thus, there are strong reasons to believe that the riboflavin operon of *A. pleuropneumoniae* has been horizontally transferred

***Mikhail S. Gelfand**
misha@imb.imb.ac.ru
***Andrey A. Mironov**
mironov@genetika.ru
***Jurges Jomantas**
jomantas@genetika.ru
***Yuri I. Kozlov**
kozlov@genetika.ru
°Danila A. Perumov
perumov@bird.macro.ru

State Center of Biotechnology GosNIIGenetika, Moscow, 1-j Dorozhny proezd, 1, Moscow, 113545, Russia.
*Ajinomoto-Genetika Research Institute, Moscow, 113545, Russia.
°St Petersburg Nuclear Physics Institute, Gatchnia, Leningrad District, 188350, Russia.

FIGURE 2. Alignment of the *RFN* element sequences

```
      0′        1′        1″      2′   add.   2″       3′    3″          4′        4″       5′       5″        0″
   ---====>    -==>     <==-    ===>  ><    <===      =>    <=         ====>     <====    ==>     <==      <====---
ba TATCCTTCgggg-cTGGGtgaaaatCCCGaccgGCGGt 23 agcCCGTgac--  8   4   8 -----tggaTTCAGtgaaaagCTGAAgccgaCAGtgaaagtCTGgat-gggaGAAGGATG
bs gtaTCTTCgggg-caGGGtggaaatCCCGaccgGCGGt 21 agcCCGTgac--  8   4   8 -----tggaTTCAGtttaa-gCTGAAgccgaCAGtgaaagtCTGgat-gggaGAAGGatg
bs caATCTTCgggg-cAGGGtgaaattCCCCTaccgGCGGt 18 agcCCGCga---  5   4   5 -----aggaTTCGGtgagattCCGGAgccgaCAGta-cagtCTGgat-gggaGAAGATgg
ca tgTTCTTCaggg-aTGGGtgaaatTCCCAatcgGCGGt  2 agcCCGCaa---  4   2   4 ------agaTCCGGTtaaactCCGGGgccgaCAGttaagtCTGgat-gaaaGAAGAAat
ca tgATCTTCaggg-cAGGGtgaaatTCCCTaccgGCGGt  2 agcCCGCgag--  3   4   3 ----tatgaTCCGGtttgattCCGGAgccgaCAGta-aagtCTGgat-gaaaGAAGATat
ef tcGTCTTCaggggcaGGGtgtaatCCCCgaccgGTGGt  3 agtCCACgac--  5   3   5 -----ttgaATTGGtgtaattCCAATaccgaCAGta-tagtCTGgat--aaaGAAGATag
pn ctaTCTTCaggg-cAGGGtgaaatTCCCTaccgGTGGt  2 agcCCACga---  3   4   3 -----atgaTTTGGtgaaattCCAAAgccgaCAGta-tagtCTGgat-gaaaGAAGATAa
pn ........................................ agtCCGTg----  3   4   3 -------gaTGTGGtgagattCCACAaccgaCAGta-tagtCTGgat-gggaGAAGAcac
py gtGTCTTCaggg-caGGGtgtgatTCCCgaccgGCGGt 14 agtCCGCg----  3   4   3 -------gaTGTGGtgtaactCCACAaccgaCAGta-tagtCTGgat-gagaGAAGACcg
tm ACGCTCTCgggg-caGGGtggaattCCCCgaccgGCGGt  3 agcCCGCg----  7   4   7 -------gaCCCGGtggaattCCGGGgccgaCGGtgaaagtCCGgat-gggaGAGAGCGT
tn TCGCTCTCgggg-caGGGtggaattCCCCgaccgGCGGt  3 agcCCGCg----  7   4   7 -------gaCCCGGtggaattCCGGGgccgaCGGtgaaagtCCGgat-gggaGAGAGCGA
dr CCTCTTTCgggg-cGGGGcgaaatTCCCCaccgGCGGt 15 agcCCGCgaa--  8  12   9 -----ccgaTGCCGcgcaactCGGCAgccgaCGGtcacagtCCGgac-gaaaGAAGGAGG
br TTGTTCTCgggg-cGGGGtgaaactCCCCaccgGCGGt 17 agcCCGCgagcg 10  15  10 gtcagcagaTCCGGtgagatgCCGGAgccgaCGGttaagtCCGgat-ggaaGAGAGCGA
bp ACGTCTTCaggg-cGGGGtgcaattCCCCaccgGCGGt 18 agcCCGCgagcg 10   4  10 gtcagcagaCCTGGtgagatgCCAGGgccgaCGGtcatagtCCGgat-gagaGAAGATGT
pa acGTTCTCaggg-cGGGGtgaaagtCCCCaccgGCGGt 19 agcCCGCgagcg 19   4  17 gtcagcagaCCCGGtgcgattCCGGGgccgaCGGtcatagtCCGgataaagaGAGAACGg
vc aTATTCTCaggg-cGGGGcgaaatTCCCCaccgGTGGt 13 agcCCACgagcg  5   4   5 gtcagcagaTCTGGtgagaagCCAGGgccgaCGGttacagtCCGgat-gagaGAGAATGa
ec TTATTCTCaggg-cGGGGcgaaatTCCCCaccgGCGGt 17 agcCCGCgagcg  8   4   8 gacagcagaTCCGGtgtaattCCGGGgccgaCGGttagagtCCGgat-gggaGAGAGTAA
st TTATTCTCaggg-cGGGGcgaaatTCCCCaccgGCGGt 67 agcCCGCgagcg  8   4   8 gtcagcagaTCCGGtgtaattCCGGGgccgaCGGttaaagtCCGgat-gggaGAGGGTAA
yp TTATTCTCaggg-cGGGGtgaaagTCCCCaccgGCGGt 40 agcCCGCgagcg 16   6  16 gtcagcagaCCCGGtgtaattCCGGGgccgaCGGttatagtCCGgat-gggaGAGAGTAA
ps cttATCTCaggg-cGGGGtggaagtCCCCaccgGCGGt 17 agcCCGCgagcg  7   9   7 gtcagcagaTCAGG.....................................................
hi gcATTCTCaggg-cAGGGtgaaatTCCCTaccgGTGGt  2 agcCCACgagcg 26   9  30 gtcagcagaTTTGGtgaaattCCAAAgccgaCAGta-aagtCTGgat-gaaaGAGAATaa
aa gcATTCTCaggg-cAGGGtgaaagtCCCCTaccgGTGGt 25 agcCCACgagcg 16   4  27 gtcagcagaTTTGGtgcgaatCCAAAgccgaCAGtgacagtCTGgat-gaaaGAGAATaa
ap taaTCTTCaggg-caGGGtgaaatTCCCgatcgGCGGt  3 agtCCGCga---  7   7   7 -----aggaACCGGtgagattCCGGTaccgaCAGta-tagtCTGgat-ggaaGAAGAaat
                                                                                                    trends in Genetics
```

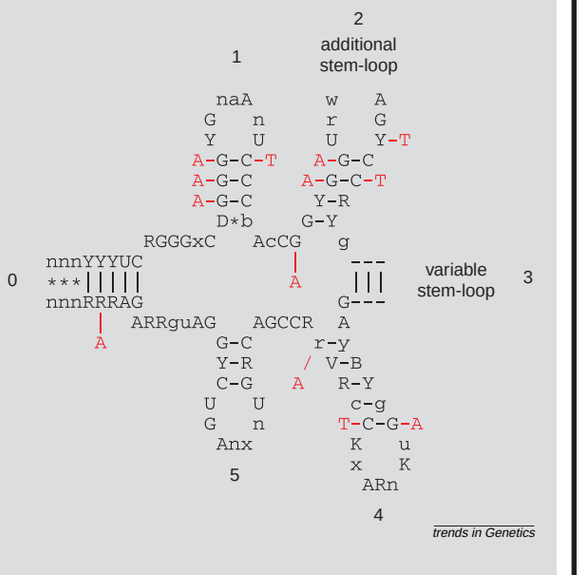
Capitals indicate base-paired positions. Red text indicates conserved positions. Green text indicates degenerate conserved positions. Black text indicates non-conserved positions. Blue text indicates non-consensus nucleotides in conserved positions. Top line: stems (double dash (=), obligatory base-pairs; single dash (-), facultative base-pair). Genome abbreviations in the left column: ba, *Bacillus amyloliquefaciens*; bs, *Bacillus subtilis*; ca, *Clostridium acetobutylicum*; ef, *Enterococcus faecalis*; pn, *Streptococcus pneumoniae*; py, *Streptococcus pyogenes*; tm, *Thermotoga maritima*; tn, *Thermotoga neapolitana*; dr, *Deinococcus radiodurans*; ba, *Brucella abortus*; bp, *Bordetella pertussis*; pa, *Pseudomonas aeruginosa*; vc, *Vibrio cholerae*; st, *Salmonella typhi*; ec, *Escherichia coli*; yp, *Yersinia pestis*; hi, *Haemophilus influenzae*; aa, *Actinobacillus actinomycetemcomitans*; ap, *Actinobacillus pleuropneumoniae*. White boxes in left column indicate *ypaA* genes.

from some Gram-positive genome. Indeed, after multiple alignment of all five protein families involved in riboflavin synthesis and construction of phylogenetic trees based on these alignments using CLUSTALX (Ref. 16), we have observed that *A. pleuropneumoniae* always clusters within Gram-positive bacteria.

The conservation of the *RFN* element, its location (when known) within the transcribed region, and the fact that mutations in the *RFN* element influence the expression rate of riboflavin genes in *Bacillus* spp. make it very likely that *RFN* is an RNA-regulatory element. It has three properties that distinguish it from other prokaryote RNA-regulatory elements. First, it is very strongly conserved both in structure and in sequence. Second, it is connected with different genes in different genomes, although all these genes are related to the riboflavin synthesis. Third, it is subject to horizontal transfer. If this is true, then it is likely that gene *ypaA* of *B. subtilis* and its orthologs in other genomes also encode a protein somehow related to the riboflavin synthesis.

It is known that various nucleotides, such as flavin mononucleotide, riboflavin, or β-nicotinamide mononucleotide can specifically bind to RNA aptamers (Refs 17–19 and references therein). Thus, although the mechanism of regulation by *RFN* is unclear, an intriguing possibility is that in this case the direct binding of the reaction product to the RNA structure is involved. It is noteworthy that mutations that lead to the overproduction of riboflavin in *B. subtilis* (Ref. 13) either change nucleotides at invariant positions or destroy conserved base-pairings (Fig. 3). Thus, we suggest that the entire structure of the *RFN* element is important for regulation.

FIGURE 3. The conserved structure of the *RFN* element



Capitals indicate invariant or absolutely conserved positions. Lower case indicates strongly conserved positions (at most two exceptions in related genomes). Red text indicates mutations leading to overproduction of riboflavin in *Bacillus subtilis* (Ref. 13). Dashes indicate obligatory base pairs. Stars indicate facultative base pairs. n: any nucleotide. x indicates any nucleotide or deletion. Degenerate positions: R = A or G; Y = C or U; W = A or U; K = G or U; B = not A; D = not C; H = not G; V = not U.

Acknowledgements
This work was partially supported by grants from the Russian State 'Human Genome' program and the Russian Fund of Basic Research under grant 99-04-48347. Preliminary sequence data were obtained from The Institute for Genomic Research website at http://www.tigr.org.

References
1 Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
2 Mironov, A.A. *et al.* (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* 27, 2981–2989
3 Vitreschak, A. *et al.* Computer analysis of regulatory patterns in complete bacterial genomes. Translation initiation of the ribosomal protein operons. *Biophysics* (in press)
4 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5849–5856
5 Winker, S. *et al.* (1990) Structure detection through automated covariance search. *Comput. Appl. Biosci.* 6, 365–371
6 Gutell, R. *et al.* (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.* 20, 5785–5795
7 Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.* 22, 2079–2088
8 Le, S-Y. and Zuker, M. (1990) Common structures of the 5′ non-coding RNA in enteroviruses and rhinoviruses. Thermodynamical stability and statistical significance. *J. Mol. Biol.* 216, 729–741
9 Le, S.Y. *et al.* (1995) Unusual folding regions and ribosome landing pad within hepatitis C virus and pestivirus RNAs. *Gene* 154, 137–143
10 Dandekar, T. *et al.* (1999) Systematic genomic screening and analysis of mRNA in untranslated regions and mRNA precursors: combining experimental and computational approaches. *Bioinformatics* 14, 271–278
11 Keener, J. and Nomura, M. (1996) In *E. coli* and *Salmonella* (Neidhardt, F.C., ed.), pp. 1417–1431, ASM Press
12 Gusarov, I.I. *et al.* (1997) Primary structure and functional activity of the *Bacillus subtilis* ribC gene. *Mol. Biol.* 31, 446–453
13 Kil, Y.V. *et al.* (1992) Riboflavin operon of *Bacillus subtilis*: unusual symmetric arrangement of the regulatory region. *Mol. Gen. Genet.* 233, 483–486
14 Fuller, T.E. and Mulks, M.H. (1995) Characterization of *Actinobacillus pleuropneumoniae* riboflavin biosynthesis genes. *J. Bacteriol.* 177, 7265–7270
15 Hemmen, F. *et al.* (1995) Cloning and sequence analysis of a newly identified *Brucella abortus* gene and serological evaluation of the 17-kilodalton antigen that it encodes. *Clin. Diagn. Lab. Immunol.* 2, 263–267
16 Jeanmougin, F. *et al.* (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403–405
17 Lauhon, C.T. and Szostak, J.W. (1995) RNA aptamers that bind flavin and nicotinamide redox cofactors. *J. Am. Chem. Soc.* 117, 1246–1257
18 Araki, M. *et al.* (1998) Allosteric regulation of a ribozyme activity through ligand-induced conformational change. *Nucleic Acids Res.* 26, 3379–3384
19 Patel, D.J. *et al.* (1997) Structure, recognition and adaptive binding in RNA aptamer complexes. *J. Mol. Biol.* 272, 645–664

The molecular clock is gone but not forgotten

What about Molecular Clocks? Symposium at the Annual Meeting of the Society for the Study of Evolution (SSE), Madison, Wisconsin, USA, 22–26 June 1999

Rasmus Nielsen
rnielsen@oeb.harvard.edu

Department of Organismic and Evolutionary Biology, Harvard University, 288 Biology Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA.

The notion of a molecular clock arose in the sixties as a product of empirical observations of apparent constancy in the rate of protein evolution. The basic idea was that substitutions occur at an approximately constant rate in DNA sequences and can, therefore, be used to date evolutionary events. The theoretical explanation for this observation was provided by Kimura's 'neutral theory', which states that substitutions occur purely as a consequence of mutation and genetic drift. Thus, if spontaneous mutations occur at a constant rate, substitutions between species will also occur at a constant rate. Much debate was generated in the seventies and eighties regarding the validity of the molecular clock. It was noted that some organisms seem to have a higher rate of substitutions than other organisms, for example, rodents appear to evolve much faster at the molecular level than primates[1].

As was evident from the symposium on the molecular clock at the Evolution Meeting in Madison, the emphasis of the discussion has shifted. With the explosion in the availability of comparative molecular data, ample evidence against a global constant molecular clock has accumulated. From testing the clock, the focus has shifted to answering the fundamental question: what are the causes for deviations from a molecular clock? Clearly there can be differences in the biological mutation rate between very divergent organisms and differences in generation time also seem to have some effect, but the main question appears to be: is darwinian selection involved in creating deviations from a constant molecular clock?

An answer to this question came from one of the speakers, Tomoko Ohta (National Institute of Genetics, Japan), the inventor of the theory of 'nearly neutral' evolution. It appears that nonsynonymous (replacement) substitutions are more irregular than synonymous (silent) substitutions, at least in some organisms. The rate of synonymous substitutions can vary strongly between divergent evolutionary lineages, presumably owing to variations in the mutational process or in the generation time between organisms. However, within evolutionary lineages, selection seems to act to make the nonsynonymous substitution process more erratic. Although this pattern might not currently be supported by all the available data, Ohta's analysis suggests that selection is of importance in creating deviations from the molecular clock.

 PII: S0168-9525(99)01858-2