Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Genome Analysis

Phylogenomics and bioinformatics of SARS-CoV

Pietro Liò[1,2] and Nick Goldman[1]

[1]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK CB10 1SD
[2]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, UK CB3 0FD

Tracing the history of molecular changes in coronaviruses using phylogenetic methods can provide powerful insights into the patterns of modification to sequences that underlie alteration to selective pressure and molecular function in the SARS-CoV (severe acute respiratory syndrome coronavirus) genome. The topology and branch lengths of the phylogenetic relationships among the family *Coronaviridae*, including SARS-CoV, have been estimated using the replicase polyprotein. The spike protein fragments S1 (involved in receptor-binding) and S2 (involved in membrane fusion) have been found to have different mutation rates. Fragment S1 can be further divided into two regions (S1A, which comprises approximately the first 400 nucleotides, and S1B, comprising the next 280) that also show different rates of mutation. The phylogeny presented on the basis of S1B shows that SARS-CoV is closely related to MHV (murine hepatitis virus), which is known to bind the murine receptor CEACAM1. The predicted structure, accessibility and mutation rate of the S1B region is also presented. Because anti-SARS drugs based on S2 heptads have short half-lives and are difficult to manufacture, our findings suggest that the S1B region might be of interest for anti-SARS drug discovery.

Can phylogeneticists and bioinformaticians help virologists to tackle SARS-CoV (severe acute respiratory syndrome coronavirus)? Phylogenetic methodology has progressed almost beyond recognition in the past decade and the study of phylogenetic relationships among species is now a valuable source of information in a variety of biological fields. The widespread use of a reliable statistical formalisation in phylogenetic and bioinformatic studies is necessary to extract the maximum information from sequence data [1].

Models of virus evolution

Recent work, although reporting impressive insights into the mechanisms of pathogenesis of SARS-CoV, has assessed the phylogeny of proteins of this virus using overly simple algorithms [2,3]. The distance methods used to assess SARS-CoV phylogeny present several disadvantages. First, by converting a sequence alignment to pairwise distances we necessarily lose some of the evolutionary information contained within the analysed sequences [1]. Second, distance methods are known to compromise the accuracy of estimates of evolutionary divergence, which are fundamental to understanding the rate and mode of viral evolution. Third, there are no known methods to test evolutionary models and estimated trees produced using the pairwise distance methodology [1].

Here we use state-of-the-art phylogenetic methods to analyse all the available coronaviridae and SARs-CoV sequence datasets to gain an insight into the origin and evolution of SARS-CoV and to narrow down the list of potential regions of its genome that might be interesting targets for drug design.

The statistically most robust method that can be used to achieve these aims is to consider the phylogenetic inference problem in a likelihood framework, using a valid model of evolution for viral genomes [1]. The choice of such a model for single-stranded RNA virus genomes is difficult. A parametric model, based on chemical or biological properties of RNA, might underestimate or completely miss important unknown constraints; for example, packaging of single-stranded RNA genomes that requires interaction with coat proteins [4]. An alternative approach is to use empirical models that are generated through comparisons of observed sequences; for example, simply counting apparent replacements between closely related sequences. Given that sequence databases are biased towards mammalian and bacterial DNA sequences, there are relatively few coding single-stranded RNA sequences to be aligned, and non-coding RNA sequences, such as rRNA or tRNA, might be subjected to different selective (e.g. structural) constraints. Because the proteins encoded by RNA genes might be subjected to functional constraints in a similar manner to non-viral proteins, it might be better to use empirical amino acid substitution models that describe the probability of fixation of amino acid changes rather than RNA models. Furthermore, relative to primary structure, the secondary structure of homologous proteins persists long after any statistically significant sequence similarity has vanished; sequences with 25% amino acid identity probably have the same secondary structure [5]. Amino acid models of evolution that incorporate protein structural information perform better than simple amino acid models [6]. Moreover, selection pressures act on protein function, which in turn is closely related to structure. Therefore, incorporating structure information into evolutionary analysis can assist in incorporating selective constraints. Here, the programme Passml-TM, which implements protein secondary structure-based models of evolution, has been used for analyses [7,8]. The first undertaking is the determination and rooting of the coronavirus phylogeny, that is, the putative origin of the sequences of interest.

Corresponding author: Pietro Liò (pietrol@ebi.ac.uk).

Rooting the *Coronaviridae* and SARS-CoV phylogenies

A large fraction (~70%) of the SARS-CoV genome encodes a replicase polyprotein, which has significant sequence homology to all of the replicase polyproteins that have been sequenced to date from the order *Nidovirales* (comprising the *Coronavirus*, *Torovirus*, *Okavirus* and *Arterivirus* genera). Therefore, this protein is a good choice for investigating the phylogenetic relationships among the family *Coronaviridae*. Viral sequences have high mutation rates and, consequently, alignments are usually difficult to prepare. Here, ClustalW is applied using standard parameters [9], followed by careful refinement of the alignments both by eye and by using the protein secondary structural information for each nidovirus sequence as predicted by PHD (http://cubic.bioc.columbia.edu/predictprotein/) [10] and PSI-PRED (http://bioinf.cs.ucl.ac.uk/psipred) [11]. Figure 1 shows the maximum likelihood tree produced using a set of homologous replicases from five SARS-CoV strains, 12 other coronaviruses representing both groups 1 and 2 of the genus [2,3], one torovirus (Breda virus) and one okavirus [yellow head (YH) virus], which were determined to most closely represent the consensus coronavirus sequence by a PSI-Blast search [12]. The coronavirus sequences allow the determination of the root of the SARS-CoV strains, whereas the torovirus and okavirus sequences provide insights into the rooting of the family *Coronaviridae* phylogeny, which is closer to group 1 than group 2 of the *Coronavirus* genus.

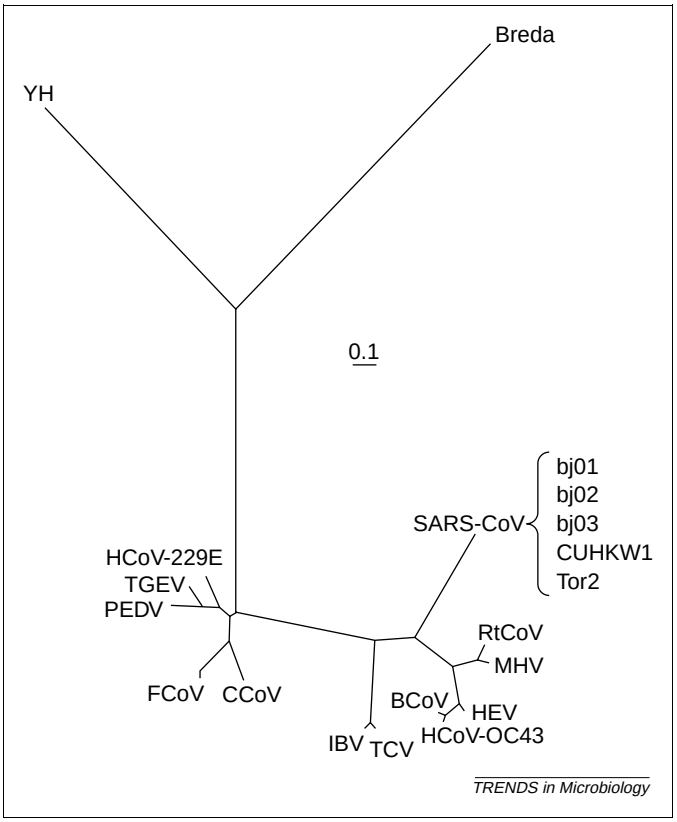

Figure 1. Maximum likelihood tree produced using Passml-TM [7] based on the replicase proteins from members of of the order *Nidovirales*, comprising 17 coronaviruses [from group 1: human coronavirus 229E (HCoV-229E), porcine epidemic diarrhea virus (PEDV), porcine transmissible gastroenteritis virus (TGEV), canine coronavirus (CCoV) and feline coronavirus (FCoV); from group 2: bovine coronavirus (BCoV), rat coronavirus (RtCoV), murine hepatitis virus (MHV), human coronavirus OC43 (HCoV-OC43) and porcine hemagglutinating encephalomyelitis virus (HEV); intermediate between groups 1 and 2: avian infectious bronchitis virus (IBV) and turkey coronavirus (TCV); and SARS-CoVs from strains Tor2, bj01, bj02, bj03 and CUHKW1], one torovirus (Breda) and one okavirus (yellow head virus, YH). The scale bar indicates evolutionary divergence corresponding to a mean of 0.1 amino acid replacements per site.

As found previously by Marra *et al.* [2] and Rota *et al.* [3], the root of SARS-CoV is closer to coronavirus group 2. All SARS-CoV strains are almost completely identical in sequence (~99% DNA sequence homology) and therefore it is not possible to get any meaningful phylogeny within the SARS-CoV group. The avian infectious bronchitis virus (IBV), which causes respiratory disease in chickens, and the turkey virus (TCV), which causes enteric disease, are clustered together; their ancestor divides groups 1 and 2 of the mammal-infecting coronaviruses. The close clustering of the chicken and turkey viruses suggests that the difference between enteric and respiratory tropisms might require only a few amino acid changes. Experiments have shown that just two point mutations in the spike (S) glycoprotein can change porcine transmissible gastroenteritis coronavirus (TGEV), a mostly enteric virus that can kill piglets, into a non-deadly virus that excels at the respiratory route but replicates poorly in the gut [13–15]. To infect enteric tract cells with TGEV, two different domains of the S protein of TGEV, mapping to between amino acids 522 and 744 and close to amino acid 219, are involved [13]. The first domain binds to aminopeptidase N (pAPN); many viruses use co-receptors, and it is probable that the second domain maps a co-receptor essential for the enteric tropism of TGEV [14,15].

The clustering of murine hepatitis virus (MHV) and rat coronavirus (RtCoV) might reflect the relatively close proximity in which the hosts reside and perhaps the similarity of murine and rat target receptors. Note that MHV receptors, including CEACAM1, have recently been identified [16,17]. The clustering of the human OC43 (HCoV-OC43), bovine (BCoV) and porcine haemagglutinating encephalomyelitis (HEV) coronaviruses might reflect conditions contributing to cross-infection in farming. The Breda torovirus is enteric, whereas the YH okavirus infects gill tissue in prawns. This indicates that the switch between enteric and respiratory tropism is a general characteristic of the order *Nidovirales*.

Analysis of the spike protein

Coronaviruses attach to host cells through the S glycoprotein [15,18]. This protein is translated as a large polypeptide that is subsequently cleaved into a receptor-binding peripheral subunit (S1) that remains non-covalently associated with a fusion-inducing membrane-spanning (S2) fragment [18].

Studies have shown that the entry of the porcine coronavirus TGEV into cells is mediated by the interactions of S1 with pAPN, an ectoenzyme abundantly expressed at the apical membrane of enterocytes covering the villi of the small intestine [19]. The fact that the S protein mediates the first interaction of the virus with human cells suggests that it might represent an excellent target for effective anti-SARS-CoV drugs; unfortunately, structural information is only available for the 3CLpro proteinase, which is part of the coronavirus replication complex (PDB accessions 1Q2W, 1P9U, 1P9S and

1LVO) [20]. The S protein shows relatively high sequence homology within two of the major coronavirus groups (>60% within group 1; >38% within group 2), whereas the homology between groups is lower (15–21% between groups 1 and 2; 15–21% between SARS-CoV and group 1 or 2). Comparative sequence analysis between the SARS-CoV sequences and sequences from groups 1 and 2 of the coronavirus genus reveals three regions of varying sequence conservation: from amino acid positions 1 to ~400, 401–680 and 681–1255. Interestingly, the first two regions correspond to the S1 fragment (we subsequently refer to these regions as S1A and S2A) and the third to the S2 fragment. Figure 2a shows a cartoon of the gene for the S protein. The region S1A is poorly conserved between groups 1 and 2 and SARS-CoV, whereas S1B is flanked by two very well conserved motifs making it easier to align the internal sequences. Notably, this region is homologous to a TGEV spike region that is reported to contain determinants for tissue tropism [13–15]. The region S2 is more conserved than S1A and S1B and consequently the alignment is easily determined; Passml-TM, which uses the evolutionary relationships of the sequences analysed to improve its predictions of secondary structure and accessibility, predicts a transmembrane helix at location 1196–1218 of the S protein. The maximum likelihood trees of S1B and S2 are shown in Figures 2b and 2c, respectively. The length of the trees and the distances between sequences reflect the sequence homologies according to the model of evolution, explaining why the S2 tree is much shorter than S1B tree.

The phylogenetic tree produced from the analysis of the spike S2 fragment from several coronaviruses indicates that SARS-CoV is closer to group 2 of the coronavirus genus than to group 1. Phylogenetic analysis of the S1 fragment from several coronaviruses indicates that SARS-CoV has an even closer relationship to group 2 viruses than the previous analysis suggests. Phylogenetic analysis of a 300 amino acid region at the terminus of S1, which we denote S1B, indicates that SARS-CoV belongs to group 2 and is closely related to MHV and RtCoV. The differences between these results might be a result of recombination events involving SARS-CoV or convergent evolution, or they might also be caused simply by chance; however, it is apparent that SARS-CoV is most closely related to group 2 of the coronavirus genus. Because the S1 fragment of MHV binds to CEACAM1, we suggest that the S1B region of the SARS-CoV spike might bind to a human CEACAM1 receptor instead of pAPN. Holmes and collaborators, who have identified receptors for MHV (murine CEACAM1a), HCoV-229E (human pAPN) and feline coronaviruses (feline pAPN) [16,17], are currently investigating this hypothesis (K. Holmes, pers. commun.). In support of the proposal that S1B is involved in receptor-binding is the fact that this region is homologous to a domain of the TGEV S protein, which is located between amino acids 522 and 744 and is also involved in receptor-binding [13,16].

Notably, TGEV mutants that lack sialic acid-binding activity contain single point mutations in the S protein (Cys155Phe, Met195Val, Arg196Ser, Asp208Asn or Leu209-Pro) [21,22]. Sialic acid-binding activity might help TGEV to resist detergent-like substances encountered during gastrointestinal passage and therefore facilitate infection of the intestinal epithelium [23]. We found that only Cys155 is conserved in SARS-CoV; this is in agreement with clinical findings that show that 20–50% of SARS patients present gastrointestinal symptoms [24]. The low conservation of the S1A region among coronavirus sequences suggests that once more strains of SARS-CoV or other closely related species are available it will become possible to use innovative comparative sequence analyses to examine positive selection that acts in this region [1].

Several important functional determinants have been discovered in fragment S2. It contains a cytoplasmic tail enriched in cysteine residues (1217–1236; Figure 2a); this is a common feature among coronaviruses and appears to be related to membrane fusion [25]. Several authors have discovered that SARS-CoV S2 contains two conserved regions of heptad repeats (913–1000 and 1151–1185; Figure 2a) [26,27] (see also a press release by W.R. Gallaher and R.F. Garry http://www.virology.net/sars/s2model.html). These heptads suggest that SARS-CoV uses mechanisms to gain entry to a cell that are similar to those used by human immunodeficiency virus (HIV), the virus that causes AIDS (acquired immunodeficiency syndrome), orthomyxoviruses and paramyxoviruses, and also Ebola. It is known that peptides derived from these repeat regions in HIV and the paramyxoviruses can specifically inhibit virus entry and subsequently viral replication [26]. Currently, SARS treatment is modeled after the drug known as T20 (http://www.hivmedicine.com/textbook/drugs/t20.htm). This drug is a complex peptide that is difficult to manufacture, has a short half-life in the human body and must be injected. This suggests that other regions of the genome, such as S1B, which might be of some interest for drug design, should be described. We report in Figure 2d the consensus protein structure estimate of S1B obtained using Passml-TM, PHD [10] and PSI-PRED [11]. Passml-TM also estimates the distribution of mutation rates along the protein and the sitewise mutation rate (Figure 2d).

Mutational spectra of SARS-CoV genomes

The SARS-CoV genome is AT rich (59%). Asymmetries in strand composition can reveal mutation bias (for example, cytosine deamination) or selection [28]. We found that the GC $[(G - C)/(G + C) = 0.02]$ and AT skew $[(A - T)/(A + T) = 0.037]$ in the SARS-CoV genome are smaller than those of the HIV genome (GC skew is 0.15 and AT skew is 0.23), which has a double-stranded RNA genome with a similar AT content. This suggests the existence of some selection on G and C distribution along the sequence to control the types of RNA secondary structure that form [29]. CG is the only dinucleotide statistically under-represented $\{f(CG)/(f(C)f(G) = 0.46$, where $f(CG)$ is the frequency of CG dinucleotides and significance is assessed according to Refs. [30,31]$\}$. Because this depletion is also found in the HIV genome but not in that of Tobacco mosaic virus, it might occur as a result of mutational bias in vertebrate cells.

Comparison of the mutation patterns in the SARS-CoV genome sequences from 16 patients shows that a large number (38/84) of the base substitutions detected at 84

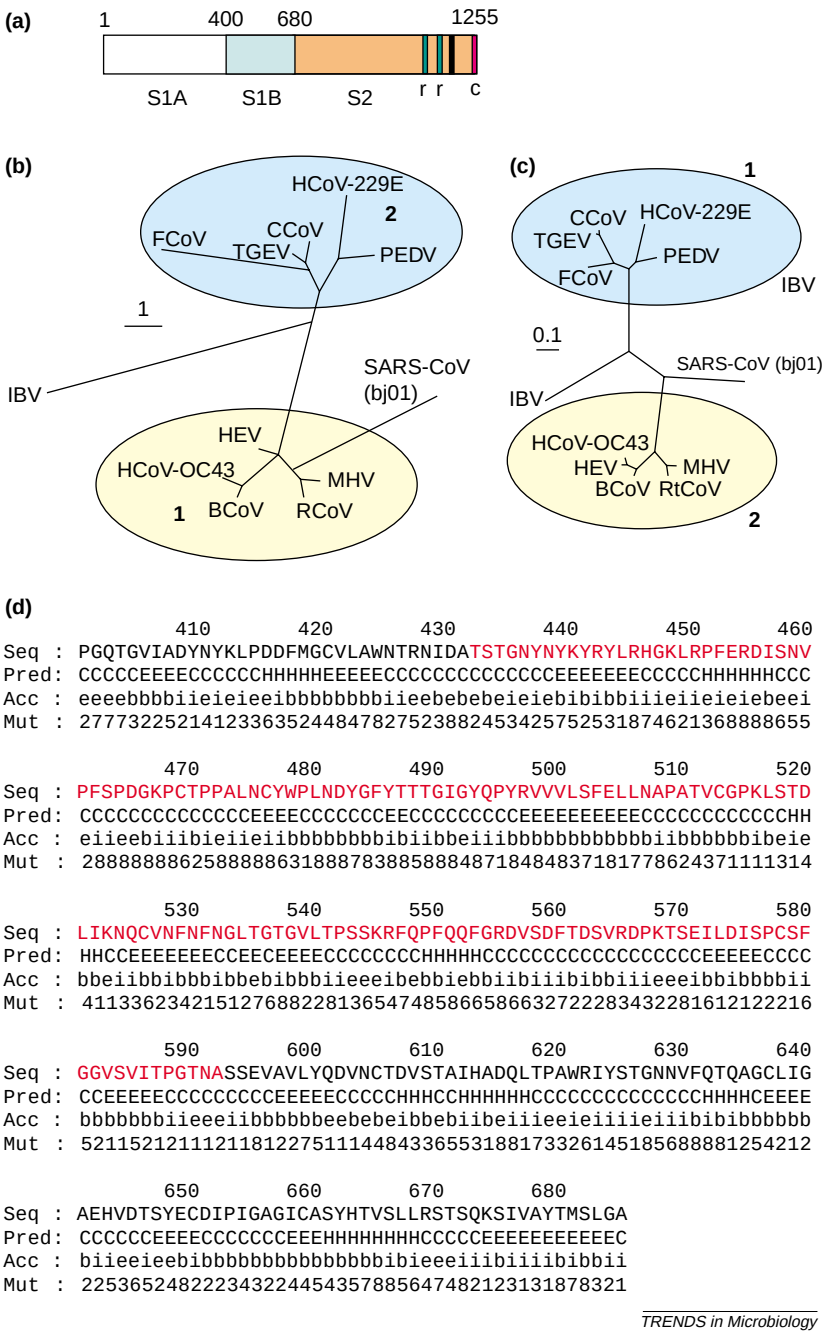

Figure 2. **(a)** Schematic diagram of the spike protein gene showing S1A, S1B (faster evolving) and S2 (conserved) regions. The inferred transmembrane region is represented in black. The heptad repeats (r) and cysteine-rich domain (c) are also shown. Maximum likelihood trees of the **(b)** S1B and **(c)** S2 regions of the spike protein. The scale bars indicate the mean numbers of amino acid replacements per site. Species used include 17 coronaviruses [from group 1: human coronavirus 229E (HCoV-229E), porcine epidemic diarrhea virus (PEDV), porcine transmissible gastroenteritis virus (TGEV), canine coronavirus (CCoV) and feline coronavirus (FCoV); from group 2: bovine coronavirus (BCoV), rat coronavirus (RtCoV), murine hepatitis virus (MHV), human coronavirus OC43 (HCoV-OC43) and porcine hemagglutinating encephalomyelitis virus (HEV); intermediate between groups 1 and 2: avian infectious bronchitis virus (IBV) and turkey coronavirus (TCV); and SARS-CoVs from strains Tor2, bj01, bj02, bj03 and CUHKW1], one torovirus (Breda) and one okavirus (yellow head virus, YH). Coloured areas indicate coronavirus groups 1 (blue) and 2 (yellow). **(d)** Predicted structure and mutation rate for the SARS-CoV S1B region. Secondary structure has been predicted (row Pred) in three classes: helix (H), sheet (E) and coil (C). Accessibility has been predicted (row Acc) in three classes: buried (b), exposed (e) and intermediate (i). Mutation rates (row Mut) are partitioned into eight classes (classes 1–8 have relative rates of evolution 0.10, 0.27, 0.44, 0.64, 0.88, 1.19, 1.65 and 2.83, respectively), inferred using the empirical Bayes method [37]. The sequence homologous to the region 522–744 of the TGEV spike protein is shown in red. Maximum likelihood trees and mutation rate analyses were computed using PassmI-TM [7].

sites occur within or near to single base and dinucleotide repeat stretches. Despite the absence of a pairing rule, the ratio of rates of transition and transversion mutations is ~2, as is often found in double-stranded DNA. The low GC and AT skews and the low number of mutations suggest that evolvability of SARS-CoV might be restricted by selective constraints acting on the RNA structure and packaging of the genome, and therefore it might also be restricted by the low fitness of its mutational neighbours. Sequence features that form stems and loops that are potentially involved in coronavirus genome packaging have been described [32,33]. Mutational neighbours with

different fitness might explain why, although some RNA viruses evolve at high rates, some RNA viruses are highly stable [34,35].

Interestingly, the ease of tropism switching as exemplified by the closeness between turkey and chicken coronaviruses (Figure 1) is favoured by the large number of viral particles in each host, by their mutation rates, by the large populations of hosts (birds and other species) and by the aerial mode of viral spread (for instance, through sneezing and faeces). These factors suggest that birds might act as powerful engines for virus evolution.

Conclusion

In this review we have highlighted that the S1 and S2 fragments of the SARS-CoV S protein have different mutational patterns. On the basis of phylogenetic evidence (Figure 2b) and the homology with the TGEV 522–744 region, it is suggested that a short region of the S1 fragment, which we denote S1B, located at positions 400–680, might be of particular interest to virologists, structural biologists and biotechnologists. The sitewise secondary structure, solvent accessibility and mutation rate of this region have been estimated; our work in progress includes further structural characterization and fold family determination. At the moment, SARS appears to be under control despite doctors having neither drugs nor a vaccine to protect against it. Because it could reappear in the future, research should proceed and hopefully our findings might assist in maintaining a feed-forward loop on SARS-CoV research between bioinformatics analysis and experimental work from microbiologists and virologists. As a final comment, it is notable that, in all the phylogenies, human coronaviruses HCoV-229 and HCoV-OC43 always cluster with porcine coronaviruses. Because Ericsson and collaborators [36] reported the identification of two homologous human proteins that act as receptors for porcine endogenous retrovirus, the benefits and risks of porcine–human xenotransplantation should be carefully balanced.

Acknowledgements

We thank Rodrigo Lopez and Ivo Cozzani for helpful suggestions. P.L. was partially supported by a BBSRC grant. N.G. is supported by a Wellcome Trust Fellowship in Basic Biomedical Research.

References

1 Whelan, S. *et al.* (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* 17, 262–272
2 Marra, M.A. *et al.* (2003) The genome sequence of the SARS-associated coronavirus. *Science* 300, 1399–1404
3 Rota, P.A. *et al.* (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300, 1394–1399
4 Butler, P.J.G. (1999) Self-assembly of tobacco mosaic virus: the role of an intermediate aggregate in generating both specificity and speed. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 537–550
5 Russell, R.B. *et al.* (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* 269, 423–439
6 Thorne, J.L. *et al.* (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.* 13, 666–673
7 Liò, P. *et al.* (1998) Combining protein secondary structure prediction and evolutionary inference. *Bioinformatics* 14, 726–733
8 Liò, P. and Goldman, N. (1999) Using protein structural information in evolutionary inference: transmembrane proteins. *Mol. Biol. Evol.* 16, 1696–1710
9 Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
10 Rost, B. (1996) Predict protein. *Methods Enzymol.* 266, 525–539
11 McGuffin, L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405
12 Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444–447
13 Ballesteros, M.L. *et al.* (1997) Two amino acid changes at the N-terminus of transmissible gastroenteritis coronavirus spike protein result in the loss of enteric tropism. *Virology* 227, 378–388
14 Godet, M. *et al.* (1994) Major receptor-binding and neutralization determinants are located within the same domain of the transmissible gastroenteritis virus (coronavirus) spike protein. *J. Virol.* 68, 8008–8016
15 Sanchez, C.M. *et al.* (1999) Targeted recombination demonstrates that the spike gene of transmissible gastroenteritis coronavirus is a determinant of its enteric tropism and virulence. *J. Virol.* 73, 7607–7618
16 Bonavia, A. *et al.* (2003) Identification of a receptor binding domain of the spike glycoprotein of human coronavirus HCoV-229E. *J. Virol.* 77, 2530–2538
17 Zelus, B.D. *et al.* (2003) Conformational changes in the spike glycoprotein of murine coronavirus are induced at 37°C either by soluble murine CEACAM1 receptors or by pH 8. *J. Virol.* 77, 830–840
18 Cavanagh, D. (1995) The coronavirus surface glycoprotein. In *The Coronaviridae* (Siddell, S.G., ed.), pp. 73–115, Plenum Press
19 Benbacer, L. *et al.* (1997) Interspecies aminopeptidase-N chimeras reveal species-specific receptor recognition by canine coronavirus, feline infectious peritonitis virus, and transmissible gastroenteritis virus. *J. Virol.* 71, 734–737
20 Anand, K. *et al.* (2003) Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science* 300, 1763–1767
21 Schwegmann-Wessels, C. *et al.* (2002) Binding of transmissible gastroenteritis coronavirus to cell surface sialoglycoproteins. *J. Virol.* 76, 6037–6043
22 Krempl, C. *et al.* (2000) Characterization of the sialic acid binding activity of transmissible gastroenteritis coronavirus by analysis of haemagglutination-deficient mutants. *J. Gen. Virol.* 81, 489–496
23 Krempl, C. *et al.* (1997) Point mutations in the S protein connect the sialic acid binding activity with the enteropathogenicity of transmissible gastroenteritis coronavirus. *J. Virol.* 71, 3285–3287
24 Zhang, J. (2003) Severe acute respiratory syndrome and its lesions in digestive system. *World J. Gastroenterol.* 9, 1135–1138
25 Chang, K.W. *et al.* (2000) Coronavirus-induced membrane fusion requires the cysteine-rich domain in the spike protein. *Virology* 269, 212–224
26 Derdeyn, C.A. *et al.* (2001) Sensitivity of human immunodeficiency virus type 1 to fusion inhibitors targeted to the gp41 first heptad repeat involves distinct regions of gp41 and is consistently modulated by gp120 interactions with the coreceptor. *J. Virol.* 75, 8605–8614
27 Kliger, Y. and Levanon, E.Y. (2003) Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy. *BMC Microbiol.* 3, 20–30
28 Frank, A.C. and Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238, 65–77
29 Huynen, M.A. *et al.* (1992) Equal G and C contents in histone genes indicate selection pressures on mRNA secondary structure. *J. Mol. Evol.* 34, 280–291
30 Karlin, S. *et al.* (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* 32, 185–225
31 Karlin, S. and Mrazek, J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* 94, 10227–10232
32 Narayanan, K. *et al.* (2003) Nucleocapsid-independent specific viral RNA packaging via viral envelope protein and viral RNA signal. *J. Virol.* 77, 2922–2927

33 Qin, L. *et al.* (2003) Identification of probable genomic packaging signal sequence from SARS-CoV genome by bioinformatics analysis. *Acta Pharmacol. Sin.* 24, 489–496
34 Burch, C.L. and Chao, L. (2000) Evolvability of an RNA virus is determined by its mutational neighbourhood. *Nature* 406, 625–628
35 Chao, L. (1997) Evolution of sex and the molecular clock in RNA viruses. *Gene* 205, 301–308
36 Ericsson, T.A. *et al.* (2003) Identification of receptors for pig endogenous retrovirus. *Proc. Natl. Acad. Sci. U. S. A.* 100, 6759–6764
37 Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936

0966-842X/$ - see front matter
doi:10.1016/j.tim.2004.01.005

Microbial Genomics

Champions of versatility

Karen E. Nelson and Claire M. Fraser

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Those of us who make our living studying microorganisms often take their metabolic versatility for granted. However, we are reminded of this versatility yet again with the recent publication by Larimer *et al.* [1] on the complete genome sequence of *Rhodopseudomonas palustris*, a purple photosynthetic bacterium that is a member of the alpha group of proteobacteria. *R. palustris* was used in some of the earliest studies that characterized aromatic degradation in microbial species and is found in very diverse environments, an observation that reflects its spectacular metabolic capabilities and ability to grow using any one of four different modes of metabolism: photosynthetic, photoheterotrophic, chemoheterotrophic or chemoautotrophic. The ability of this organism to adjust its metabolism so dramatically in response to changes in energy and nutrient availability sets *R. palustris* apart from many of its close relatives.

The *R. palustris* genome is composed of a single circular chromosome (5.46 Mb) and one 8.4 kb plasmid and encodes 4,836 genes. The relatively large size of the *R. palustris* genome is reminiscent of other ubiquitous environmental bacteria for which we have complete genome sequences, including *Streptomyces coelicolor* (8.7 Mb), *Streptomyces avermitilis* (9.0 Mb) and *Pseudomonas putida* (6.6 Mb). The *Streptomycete* genomes contain an unprecedented number of gene clusters that encode known or predicted proteins involved in secondary metabolism. These gene clusters might represent DNA that was acquired through lateral gene transfer at some point in the evolution of the *Streptomycetes*. By contrast, *R. palustris* has a minimal number of insertion sequence elements, and few regions of atypical nucleotide composition which, taken together, suggest that lateral gene transfer has not played a major role in shaping this genome. The genes that encode components of major metabolic pathways are also randomly distributed throughout the *R. palustris* genome.

In *R. palustris,* key genes that code for bacteriochlorophyll and carotenoid biosynthesis, as well as those for membrane-bound reaction centre complexes of photosynthesis were found to be clustered in a 55 kb region of the genome. Forms I and II of RubisCO, the key enzyme of the Calvin–Benson–Bassham pathway of carbon dioxide fixation are encoded by the genome, as well as two RubisCO-like proteins (so far only identified in the photosynthetic *Chlorobium tepidum*) [2]. Based on the biochemical characterization of the RubisCO-like protein in *C. tepidum* [3], in *R. palustris* these proteins probably play a role in sulfur metabolism.

Approximately 31% of the predicted coding sequences in the genome are devoted to energy metabolism. For example, many predicted protein sequences for the oxidation of inorganic compounds including thiosulfate and hydrogen, which provide reducing energy for carbon dioxide and nitrogen fixation are encoded, and also carbon monoxide and formate dehydrogenases. Analysis of the genome also highlights the ability of *R. palustris* to degrade heterocyclic aromatics and chlorinated benzoates, and reveals the presence of four-ring cleavage pathways.

As with the soil bacteria for which we have complete genome sequences, just over 9% of the *R. palustris* genome is devoted to regulation. This is probably a reflection of the need of these species to sense variations in environmental conditions and respond accordingly. This diversity is also reflected in the high percentage of the genome that is devoted to transport capabilities and chemotaxis. As ~30% of the *R. palustris* genome consists of genes that are either hypothetical or conserved hypothetical proteins, or of unknown function, the biochemical characterization of these predicted protein sequences will undoubtedly reveal additional new physiological capabilities for this species.

R. palustris has been studied for its metabolic potential for close to forty years [4], however, the genome sequence revealed a greater set of predicted coding sequences for the biodegradation and metabolism of aromatic compounds than had been expected, as well as an ability to combine these pathways under anaerobic and aerobic conditions. It is anticipated that, as with the genomes of other metabolically diverse environmental species such as *Deinococcus radiodurans*, the genome of *R. palustris* can serve as a model to increase our understanding of how organisms can use diverse metabolic capabilities to respond to changes in substrate, light and other environmental cues.

Corresponding author: Claire M. Fraser (cmfraser@tigr.org).