8. Nair, M. M., *Mem. Geol. Soc. India*, 1987, **29**, 450–458.
9. Ramasamy, Sm., Panchanathan, S. and Palanivel, R., Proceedings of International Geoscience and Remote Sensing Symposium, Michigan, 1987, pp. 1157–1161.
10. Subrahmanya, K. R., *Curr. Sci.*, 1994, **67**, 527–530.
11. Moody, J. D. and Hill, M., *J. Bull. Geol. Soc. Am.*, 1956, **67**, 1207–1246.
12. Lepichon, X., *J. Geophys. Res.*, 1968, **73**, 3661–3697.
13. Radhakrishna, B. P., in Proceedings of the Seminar on Geomorphological Studies in India, Univ. of Saugar, Sagar, 1967.
14. Ghosh, B. N. and Zutshi, P. L., *Geol. Surv. Ind. Spl. Publ. No. 24*, 1987, pp. 309–318.

Sm. RAMASAMY

Centre for Remote Sensing,
School of Earth Sciences,
Bharathidasan University,
Tiruchirapalli 620 023, India

Interpretation of nucleotide/protein sequence data: Some pitfalls

Identification of unknown genes and their products by computer search have received high priority among molecular biologists. However, since biologists are not always computer analysts at the same time, there always exists a finite probability of making errors in analysing sequences and in interpretation of data which can easily be avoided with some advice from computer analysts. We would like to highlight some of the precautionary measures that need to be taken to avoid the inherent pitfalls in sequence data and their proper interpretations.

The worst and a very common source of errors in DNA sequences is the contamination of vector sequences arising during cloning and subcloning of DNA fragments. In 20,000 sequences in the GenBank 63 alone, more than 50 instances of cloning vector contamination have been detected[1]. In a smaller number of cases the anomalous sequences might have arisen during editing, but in a majority of the cases, large blocks of vector sequences contaminated the actual sequence. Incorporation of anomalous sequences, particularly in the coding regions, not only hinders the recognition of low-level homologies and consensus sequences, but can also exhibit false similarities with other sequences, leading to erroneous conclusions. An example of such an error is in the nucleotide sequence of the *recA* gene of *Vibrio cholerae*[2] published from our laboratory. While careful sequencing of both strands and comparison of the sequence-based restriction maps with the map of the cloned DNA fragments might eliminate the problem, vector contamination can be easily and quickly detected and removed using computer programs that are now available[1]. The contamination can also be removed by comparing the sequence with the dataset of vector sequences available at the GenBank.

'Frameshift' is another frequently occurring error in DNA sequences[3,4]. This happens when a base is either missed or added during reading of sequencing gels. Posfai and Roberts[5] have detected many such errors in EMBL (release 24) and GenBank (release 56) using the program DETECT, which examines alternative reading frames from related proteins. The program BLASTX[4,6] based on BLAST[7] algorithm can also predict frameshift errors by comparing the translated nucleotide sequences from all six reading frames with a protein database. When the similarity to a protein switches from one frame to another in the same strand of the query sequence, there is definitely a frameshift error in the sequence[4]. However, with the simultaneous translation and alignment algorithms, proteins with >30% sequence identity can be reliably recognized even in the presence of 1% frameshifting error rates and 5% base substitution rates[8].

In nucleotide sequence analysis, unknown complete or partial open reading frames (ORFs) are often encountered along with the sequence of the gene of interest[4,6]. These ORFs, overlooked in most cases by the investigators, might represent useful functions. Two computational approaches are now used either independently or in combination to determine whether the overlooked putative ORFs are indeed bonafide genes[3,4,6]. The first is an intrinsic approach, which distinguishes the coding regions from the non-coding ones on the basis of the statistical analysis of some parameters of the sequence without referral to any other sequence[4,6,9,10]. The software that has been extensively used for this purpose in human genome sequence analysis[2] is GRAIL[11], which utilizes the sensor-neural network approach to evaluate the intrinsic properties of the DNA sequence like frame bias, dinucleotide fractal occurrence, etc. This approach can also be used in the analysis of small genome after proper modification. Using the GenMark method based on phased Markov Chain model, Borodovsky *et al.* predicted expressed ORFs in the unannoted regions of *Escherichia coli* genomic DNA from EcoSeq6 database[12]. The second approach for predicting genes is extrinsic and involves comparison of the putative deduced amino-acid sequence with protein sequence databases and searching for motifs[4,6]. If the deduced amino-acid sequence of the putative ORF shows 'significant' similarity to one or more proteins in the database, it is almost certain that the putative ORF represents a bonafide gene. It may be pseudo- or cryptic gene, but is definitely not a part of non-coding region[4]. Using these two approaches, a large number of new genes in the unannoted regions of *E. coli* genome have been identified and searches for genes that have escaped detection so far in several other organisms are in progress.

To search for homology of a newly generated protein sequence with other sequences, i.e. databank one or the other alignment programs are used[7,13–15]. If after suitable gaping more than 25% residues of two protein sequences longer than 100 residues are identical, they must be evolutionarily and/or functionally related[16]. If the identity is between 15 and 25%, the sequences are considered to be marginally similar and in the language of computational biology, this range is known as 'twilight zone'[16]. Although this criterion of assessing 'similarity' between two sequences is acceptable, a more rigorous approach is provided by the BLAST program[7] where a probability factor (p) is assigned to each alignment, which decides whether the observed similarity between two sequences could have been obtained by chance[4]. Usually $p < 10^{-4}$ suggests genuine homology between two sequences and $10^{-4} < p < 10^{-1}$ constitutes the 'twilight zone'[4]. When the similarity between the two sequences falls in the twilight zone, some caution should be taken to decide whether the sequences are genuinely related or the twilight alignment is an artifact. Ignoring alignments with $p > 10^{-4}$ might eliminate the possibility of detecting false relatedness, but such stringent restrictions might hinder detection of weak but genuine relatedness between two sequences. What is recommended in these cases is to check whether the alignment in the twilight zone is conserved in a protein family. The software MACAW[17] can detect subtle similarities between protein sequences.

Protein sequences which have 'biased amino acid composition' often generate erroneous conclusions regarding its relatedness with other proteins with similar type of bias. For example, heat-shock proteins have clusters of charged amino-acid residues[18], the human transcription factors are rich in glutamine[19], zinc finger proteins and homeo-box regions of DNA-binding proteins have bias towards positively charged amino-acid residues[20]. If a new sequence has any of these characters, it will show similarity in the region of bias even when the proteins are functionally not related. This problem can be solved by filtering out the regions of bias before searching for similarity with other proteins. The program SEG can filter out the biased or 'low-complexity' regions from a sequence[6].

Computers function objectively and in computer searches there is no scope for prejudice that is often associated with experimental studies. This advantage of computational methods, the explosive growth in sequence databases and the availability of a wide variety of analytical software packages have, in recent years, encouraged molecular biologists to undertake computational analysis of nucleotide and protein sequences to derive a variety of information. To exploit fully the sequence data it is, however, necessary to assimilate new approaches of computational methods and this can only be realized through users having strong background not only in molecular biology but also in computers.

1. Lamperti, E. D., Kittelberger, J. M., Smith, T. F. and Villa-Komaroff, L., *Nucleic Acids Res.*, 1992, **20**, 2741–2747.
2. Ghosh, S. K., Biswas, S. K., Paul, K. and Das, J., *Nucleic Acids Res.*, 1992, **20**, 372.
3. Doolittle, R. F., *Curr. Op. Biotech.*, 1994, **5**, 24–28.
4. Koonin, E. V. and Rudd, K. E., *Trend Biochem. Sci.*, 1994, **19**, 309–313.
5. Posfai, J. and Roberts, R. J., *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 4698–4702.
6. Borodovsky, M., Rudd, K. E. and Koonin, E. V., *Nucleic Acids Res.*, 1994, **22**, 4756–4767.
7. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., *J. Mol. Biol.*, 1990, **215**, 403–410.
8. States, D. J. and Botstein, D., *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 5518–5522.
9. Staden, R., *Methods Enzymol.*, 1990, **183**, 163–180.
10. Bougueleret, L., Tekaia, F., Sauvaget, I. and Claverie, J. M., *Nucleic Acids Res.*, 1988, **16**, 1729–1738.
11. Uberbacher, E. C. and Mural, R. J., *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 11261–11265.
12. Rudd, K. E., *ASM News*, 1993, **59**, 335–341.
13. Lipman, D. J. and Pearson, W. R., *Science*, 1985, **227**, 1435–1441.
14. Smith, T. F. and Waterman, M. S., *J. Mol. Biol.*, 1981, **147**, 195–197.
15. Feng, D. F., Johnson, M. S. and Doolittle, R. F., *J. Mol. Biol.*, 1985, **21**, 112–125.
16. Doolittle, R. F., in *Of URFs and ORFs*, Univ. Science Books, Mill Valley, 1987.
17. Tatusov, R. L., Altschul, S. F. and Koonin, E. V., *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 12091–12095.
18. Karlin, S., Blaisdell, B. E. and Brendel, V., *Methods Enzymol.*, 1990, **183**, 388–402.
19. Courey, A. J. and Tijan, R., *Cell*, 1988, **55**, 887–898.
20. Brendel, V. and Karlin, S., *Proc. Natl. Acad. Sci. USA*, 1989, **86**, 5698–5702.

CHITRA DUTTA
JYOTIRMOY DAS

Biophysics Division,
Indian Institute of Chemical Biology,
Calcutta 700 032, India