Sequence banks

Searching for sequence similarities

from Temple F. Smith and Christian Burks

WITH the recent establishment of nucleic acid sequence data bases in both Europe[1] and the United States[2], highly organized and nearly complete listings of the known experimentally determined sequences are now available for exhaustive computer-assisted sequence comparisons. As several recent papers and a workshop* on the computational analysis of DNA sequences make clear, it now seems practicable to carry out global searches of the data for functional, structural and evolutionary relationships between sequences. Initial searches carried out at the National Institutes of Health and Los Alamos National Laboratory have brought to light several fascinating potential relationships, including at least one intra-sequence repeat that may represent a misadventure in cDNA cloning.

Three problems associated with such global searches were discussed at the meeting: increasing their efficiency, confirming their mathematical rigour and/or reliability and, finally, determining the statistical significance of the identified similarities. While there exists no known complete analytical solution of the last problem, the large size of the data bases provides for empirical analysis of the significance of any particular result.

The first two questions are interrelated, and two different approaches to addressing them were reported. D. Lipman and W. Wilbur (NIH) have developed a fast heuristic algorithm[3], while at Los Alamos National Laboratory a rigorous generalization[4] of the Needleman–Wunsch algorithm has been used. The former approach results in impressively fast searches on a medium-sized computer with reasonable recognition of sequence similarities. In contrast, although the Needleman– Wunsch algorithm is better understood, it requires access to a very large computer to achieve reasonable speeds (44,000 comparisons of all pairs of vertebrate sequences among themselves required approximately 170 min on a CRAY).

A number of very interesting similarities have been identified using the currently implemented local algorithm[4] on the vertebrate portion of the NIH Nucleic Acid Sequence Data Bank[2] at Los Alamos. First, all the obvious homologies among such well studied gene families as the haemoglobins are identified with similarity values from three to twenty 'standard deviations' above the mean similarity value. It should be noted, though, that standard deviations have to be derived empirically because distributions of similarity values obtained for the maximally similar local subsequences are rarely normal. However, such a simple measure of an outlier's relative significance is, no doubt, still useful.

A complete investigation of the similarities between the vertebrate sequences and the members of the Alu/B1/4.5S middle repetitive sequences reveals not only the known homologies but also examples of a B1-like sequence within a mouse immunoglobulin κ chain on the complement strand and in a mouse histone gene. A global search using such a dispersed family is instructive in that known family member similarities are often just barely distinguishable from apparent chance similarities within the large data set used. On the other hand, what appears as unquestionably significant similarities between the whitefish antifreeze protein and the spacer between the *Xenopus laevis* 5.8S and 23S rRNA and numerous other known sequences indicates only similarity between the high CC(purine)CC repeat in the fish and other very C- or G-rich regions. Weak similarities supporting previously noted homologies at the amino acid level between such 'unrelated' proteins as the chick ovalbumin and the primate α_1-antitrypsins[5] were easily identified. Similarities were also found between other unrelated sequences such as the *X. laevis* rRNAs and a herpes simplex virus transcript intron, as well as between the anglerfish preproinsulin and a mouse immunoglobulin heavy chain.

Finally, the importance of these global searches showed up most strikingly when comparisons involved the complementary strand sequence. Here several cases were found where the most similar sequence was its own complement. While this is not surprising for structural RNAs considering their secondary structure requirements, it is unusual for protein-coding regions. One case has been found just outside of the 3′ end of the human preproinsulin transcription region. This is an imperfect inverted repeat of approximately 150 bases.

Even more interesting is the perfect 113-base inverted repeat in the human leuenkephalin precursor cDNA[6]. Consideration of the position of this particular inverted repeat (the initial 113 bases and 725 bases in the coding region) suggests that it may simply be a 'hooking error' — that is, the reverse transcriptase that was used to backtranslate the RNA self-primed when it reached the end of the sequence and worked back along the sequence a short distance. Further investigation is needed, particularly since homologous sequences are known in several other organisms that could result from similar errors.

In assessing the statistical significance of potential similarities, one must consider whether the disproportionate representation of some out of all biologically realizable DNA sequences in current data bases might not impose a bias on any evaluation of significance. There is a natural tendency for biochemical research to gravitate towards species and genes for which the genetics and biochemistry are known. Though this and similar prejudices may bias the apparent centre of distribution of chance similarities, the range of the bias from comparison to comparison is small relative to the width of the 'chance distribution' itself and thus has only a slight influence on the empirical identification of significant outliers.

W. Goad and M. Kanehisa have developed a heuristic approach[7] to the quantitative estimation of the significance of the sequence similarities; the current popular approach is to use Monte Carlo simulations. Higher-level correlations, such as the known domain-dependent base nearest-neighbour frequencies (as suggested by W. Fitch[8]), or even the recognized codon usage correlations[9], should readily be included in the latter. In any case, to date, the empirical comparisons with the known biologically realizable sequence distributions appear to be very conservative means of estimating statistical significance.

It will shortly become routine for each new entry in the US nucleic acid data bank, GenBank[10], to be compared with all previous entries, and this should bring to light numerous new relationships. However, as noted by A. Maxam at the Aspen meeting, and so clearly reinforced by the recent work on various DNA-binding proteins[11,12], functional sequence similarity may not be adequately reflected in just the simple base-specific comparisons. Rather, comparative searches must be based, in addition, on equivalent local twist and chemical sites. □

Temple F. Smith and Christian Burks are in the Los Alamos National Laboratory Theoretical Biology Group, Los Alamos, New Mexico 87545.

*The workshop was held at the Aspen Center for Physics, Aspen, Colorado.

1. Walgate, R. *Nature* **296**, 596 (1982).
2. Lewin, R. *Science* **218**, 817 (1982).
3. Lipman, D. & Wilbur, W.J. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
4. Smith, T. & Waterman, M.S. *J. molec. Biol.* **14**, 195 (1981).
5. Leicht, M. *et al. Nature* **297**, 655 (1982).
6. Comb, M. *et al. Nature* **295**, 663 (1982).
7. Goad, W.B. & Kanehisa, M. *Nucleic Acids Res.* **10**, 247 (1982).
8. Fitch, W. *J. molec. Biol.* **163** (in the press).
9. Grantham, R. *et al. Nucleic Acids Res.* **9**, 43 (1981).
10. Jordan, E. *et al. Science* **218**, 108 (1982).
11. Sauer, R.T. *et al. Nature* **298**, 447 (1982).
12. Pabo, C.O. & Lewis, M. *Nature* **298**, 443 (1982).