


EUROPEAN MOLECULAR BIOLOGY ORGANISATION (E.M.B.O.)

WORKSHOP ON

PATTERN ANALYSIS IN NUCLEIC ACID

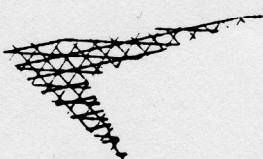

AND PROTEIN SEQUENCES

Saint-Agnan, Bourgogne (France) October 27th - 30th, 1981



THE LESSONS . One aim of the workshop is to evaluate critically the work that has been done or can be done in the exploitation of sequence data. *EVOLUTION*. Construction of phyletic trees, search for homologies, guesses on ancestral sequences. *STRUCTURE*. Prediction of secondary structures in proteins and nucleic acids, of folding domains in proteins, of melting domains in nucleic acids. *PATTERNS*. Search for repeats, palindromes, periodicities, cleavage sites, recognition signals. *CONSTRAINTS*. How the genetic code and other factors may constrain the sequences. Sequence reconstruction from fragments. Design of optimal sequences for synthesis.

THE TOOLS. The second aim of the workshop is to explore all possible ways of going further in the exploitation of the sequence data. *EXISTING TOOLS*. Efficient algorithms. Editing and presenting the data. Graphic analysis. Statistics. *ARTIFICIAL INTELLIGENCE*. The methods of pattern analysis in other fields : image enhancement (e.g., in electron microscopy) , Fourier analysis (crystallography), the methods of structural linguistics , adaptive filters and adaptive memories (perception) .



Inquiries and applications : write, before April 30th to :

*Jacques Ninio / IRBM, Tour 43 / 2 Place Jussieu /
75251 Paris cedex 05 / France*

same representation, the *E. coli* 16S RNA appears to be more homogeneous. It is not unlikely that in bacteria (or more generally, in organisms with small genomes), there are general selective constraints on DNA sequences which tend to promote statistical homogeneity all along the genes.

The last example is that of human α -hemoglobin (Fig. 5). The separation in introns and exons is obvious. The relationship between gene and pseudogene is less obvious. The positioning of the intron-exon junction suggests that the frontiers may have moved somewhat, due to nibbling of introns by adjoining exons. The introns of β -hemoglobin, shown in reference 3, seem unrelated to those of α -hemoglobin.

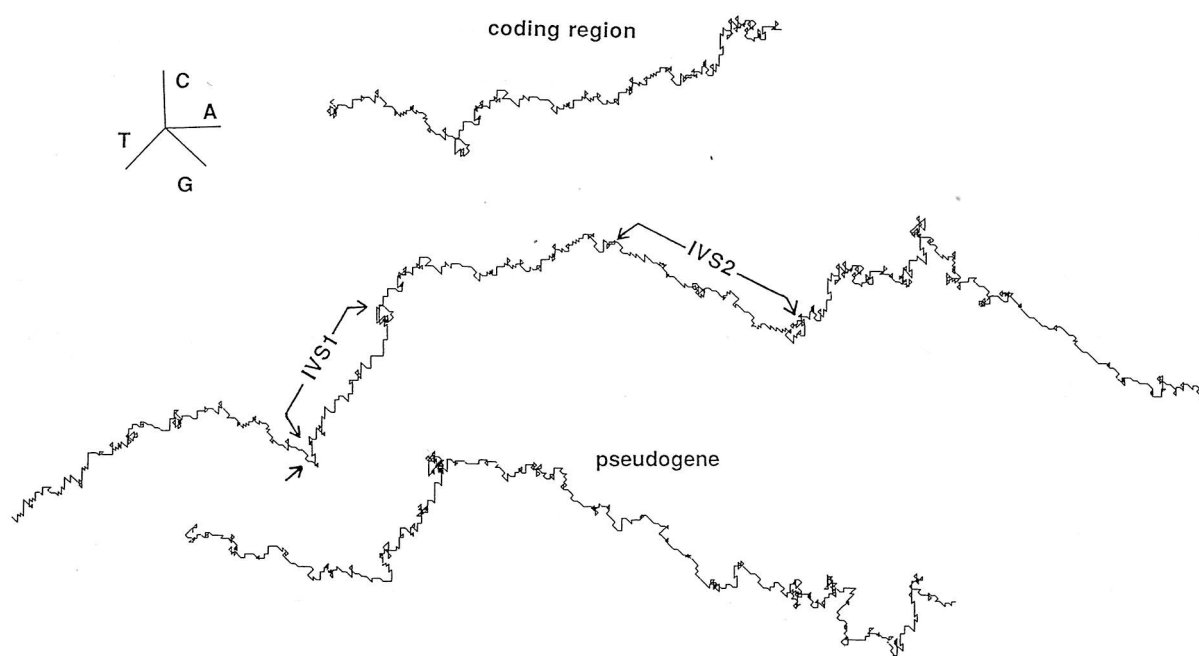


Fig. 5. Human α -hemoglobin. The sequence in the middle is that of a human α -hemoglobin gene.¹⁹ The introns are labeled IVS1 and IVS2, and the complete coding region is shown on top. The bottom sequence is that of a pseudogene for human α -hemoglobin.²⁰

One conspicuous property of vectorial representations is that they present compressed or expanded regions. Roughly, a compressed region is one in which there is a high turnover of the four nucleotides. In extended regions, there is a bias against one or more nucleotides. In all cases examined so far, the compressed/extended criterion failed to lead to biologically important insights. On the other hand, the dissection of the sequence in regions of constant local direction often reflects known biological features.

Three different correspondences between vectors and nucleotides were used for the three examples of Figs. 1–5. Twelve different codings are possible,