Genetic Engineering 2

edited by
Robert Williamson

Genetic Engineering 2

Edited by

Robert Williamson

Professor of Biochemistry,
St Mary's Hospital Medical School,
University of London

ACADEMIC PRESS · 1981

A Subsidiary of Harcourt Brace Jovanovich, Publishers

London · New York · Toronto · Sydney · San Francisco

ACADEMIC PRESS INC. (LONDON) LTD
24/28 Oval Road,
London NW1

United States Edition published by
ACADEMIC PRESS INC.
111 Fifth Avenue,
New York, New York 10003

British Library Cataloguing in Publication Data

Genetic engineering.
Vol. 2
1. Genetic engineering — Periodicals
I. Williamson, R
575.1′05 QH 442

ISBN 0–12–270302–2

LCCCN 80–41976

Printed in Great Britain at the Alden Press
Oxford London and Northampton

Contributors

J.D. Beggs *Cancer Research Campaign, Eukaryotic Molecular Genetics Research Group, Department of Biochemistry, Imperial College, London SW7 2AZ, UK*

H.H. Dahl *Laboratory of Gene Structure and Expression, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*

R.A. Flavell *Laboratory of Gene Structure and Expression, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*

F.G. Grosveld *Laboratory of Gene Structure ana Expression, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK*

A.J. Jeffreys *Genetics Department, University of Leicester, Leicester LE1 7RH, UK*

A.D.B. Malcolm *Biochemistry Department, St Mary's Hospital Medical School, University of London, London W2, UK*

Preface

In the preface to Volume 1 of this series, I commented on the revolutionary nature of genetic engineering, and the fact that it permits qualitatively different kinds of scientific results to be obtained. The article in this volume on gene evolution illustrates this: Alec Jeffreys describes the study of divergence of coding and non-coding DNA sequences for related species, and shows how fossil evidence, protein sequence data and the study of genes using recombinants all tie together to give new evolutionary concepts. Dick Flavell and his colleagues outline the way in which genomic libraries can be made and analysed, a companion article for that from Jeff Williams in Volume 1, which may prove particularly interesting to those searching for particular human genes implicated in hereditary disease. Alan Malcolm outlines various considerations for the user of restriction enzymes: as many molecular biologists have no enzymological training, and regard enzymes (correctly) primarily as commercial products of remarkably high purity and reliability that are bought from suppliers, I think it particularly apt that such a clear summary should be available pointing out that there are basic enzymological properties of restriction endonucleases whose study leads to better experimental results. Finally, Jean Beggs gives a brief outline of new vectors available which allow cloning in yeast, a host which may have great commercial usefulness, as well as representing a stepping stone between prokaryotes and higher plants and animals.

It has again been a pleasure to edit this volume: the contributions will, in my view, be of interest both to the dedicated molecular biologist and the good student (of any age) hoping to learn more about the field. My colleagues at St Mary's, and at Academic Press, have been unfailingly helpful. However, the greatest pleasure has been the number of unsolicited positive comments from colleagues complimenting Volume 1. Whether this series meets a need can only be assessed in terms of its usefulness to those working in, and studying, recombinant DNA technology.

London, 28 April 1981 *Bob Williamsor*

Contents

Recent studies of gene evolution using recombinant DNA

A.J. Jeffreys

The use of genomic libraries for the isolation and study of eukaryotic genes

H.H. Dahl, R.A. Flavell and F.G. Grosveld

The use of restriction enzymes in genetic engineering

A.D.B. Malcolm

Gene cloning in yeast

J.D. Beggs

Recent studies of gene evolution using recombinant DNA

A. J. JEFFREYS

Genetics Department, University of Leicester, Leicester, UK

I Introduction

By comparing gene sequences, we can learn something of the mechanisms by which DNA evolves. This review will survey some recent advances in our understanding of the evolution of animal genes, which have resulted from the application of recombinant DNA methods to the analysis of gene structure and function.

A Setting the scene: the molecular evolution of proteins

The foundations of the study of molecular evolution were established by comparing amino acid sequences of related proteins (Dayhoff, 1972). Homologous protein sequences in different species were found to differ in a phylogenetically consistent fashion: the more closely related the species, the more similar the amino acid sequences. By comparing sequences, detailed molecular phylogenies were derived that in general reflected the evolutionary relationship between species as deduced from taxonomic and palaeontological studies. Most significantly, individual proteins were found to change in sequence during evolution at a constant rate, irrespective of the particular lineage being studied (see Wilson *et al.*, 1977). This evolutionary molecular clock runs at very different rates for different proteins, with a 400-fold difference between the highly conserved histone H4 and the rapidly evolving fibrinopeptide B.

Amino acid substitutions do not occur at random positions within polypeptides; for example, a number of residues in haemoglobin, particularly those associated with the haem binding site, are important for function and tend to remain invariant during evolution. Similarly, the difference in evolutionary clock rates between histones and fibrinopeptides suggests that there are relatively very few histone residues which can be replaced without affecting polypeptide function.

Amino acid sequencing has also revealed the existence of families of related proteins within a single species. For example, there are clear homologies between the α-, β-, γ-, δ- and ϵ-globin polypeptides of man, sufficient to establish that the corresponding globin genes must have arisen by some process of sequential gene duplication from an ancestral globin gene. By using clock rates for globin sequence divergence, it is possible to estimate the times at which these gene duplications occurred. For example, the first duplication giving rise to the α- and β-globin genes probably occurred some 500 million years ago, consistent with the existence of α- and β-globin in all higher vertebrates but not in primitive chordates such as the lamprey (Dayhoff, 1972).

These phylogenetic analyses have been complemented by studies of genetic variation of polypeptide sequences within a species (see Harris and Hopkinson, 1972; Harris, 1980). Both polymorphic and rare variants exist at many loci, and can be regarded as a major reservoir of genetic variability, which, through natural selection or genetic drift, can lead to the fixation of new primary polypeptide sequences in a population. Much controversy exists over the mechanisms responsible for maintaining this variability. The selectionist school argues that most or all polymorphisms are maintained by selection through, for example, heterozygous advantage, whereas neutralists argue that most variants are selectively neutral and attain polymorphic frequencies or become fixed by a process of random genetic drift. Similarly, neutralists would maintain that the clock-like evolution of polypeptides is the consequence of a constant rate of appearance and fixation, by drift, of selectively neutral protein variants. The selectionist argument is that amino acid substitutions are adaptive and that the clock-like mode of evolution reflects a species' response to a constantly changing environment (see Wilson *et al.*, 1977; Gale, 1980; Harris, 1980).

B Why study molecular evolution at the DNA level?

Little of the DNA in higher eukaryotes is used for coding proteins; the function of the remaining DNA, including intervening sequences, gene spacers and satellite DNAs is enigmatic (see Orgel and Crick, 1980). Possible genetic functions for these non-coding elements and the evolution of these functions can only be studied by analysing the molecular evolution of DNA. In particular, it is possible that major changes in morphology during evolution are the result of alterations not of proteins, but of extragenic regulatory elements (King and Wilson, 1975). Clearly, the analysis of protein evolution tells only part of the story, and must be extended by studies on DNA.

C Early studies on the evolution of DNA

Before the advent of recombinant DNA technology, evolutionary studies were confined to the entire genome or to readily isolated DNA subfractions (see Lewin, 1974). The first comparisons of genome size (C value) revealed that even closely related species can have widely differing C values; the mechanisms and evolutionary significance of these rapid genome expansions and contractions are still unknown. Repetitive and unique DNA sequences from different species were compared in more detail by DNA annealing. As with proteins, DNA sequence divergence could be used to construct species phylogenies, although there is some doubt whether nuclear DNA sequences evolve in a clock-like fashion (Kohne, 1970; Bonner *et al.*, 1980). The next major advance in evolutionary studies came with detailed analyses of repetitive DNA sequences, including apparently non-coding DNA sequences (such as satellite DNAs) and coding DNA (such as histone genes and ribosomal DNA)(see Hood *et al.*, 1975; Tartof, 1975; Kedes, 1979; Long and Dawid, 1980). Many of these sequences are arranged in tandem repetitive families which have probably arisen, in part at least, by tandem gene duplication. Expansion and contraction of the size of these families by unequal crossing over between homologous repeats was postulated as a mechanism for maintaining sequence homogeneity within a family. Other mechanisms, such as transposition or chromosomal translocation, were invoked to account for the dispersal of some repetitive sequence families (Tartof, 1975).

D The new studies: evolution of single copy genes

Within the last few years recombinant DNA technology has allowed us to detect and clone single gene sequences from the DNA of higher organisms, and has opened the way to a detailed analysis of the evolution of single genes and their associated DNA sequences. Although the field is still in its infancy, major advances have already been made in analysing the organization and evolution of gene clusters and in using phylogenetic comparisons to determine rates and modes of DNA sequence divergence within and outside genes. The discovery of intervening sequences within genes has led to some fascinating speculation on the evolution of gene structure in eukaryotes. Non-functional pseudogenes have been discovered, and apparently represent the evolutionary relics of once active genes. Parallel or "concerted" evolution of gene pairs has been shown to occur as a result of DNA sequence exchange between duplicated loci. No doubt many other evolutionary surprises are awaiting us as the analysis of single copy genes continues.

E Summary

Comparisons of genes isolated by recombinant DNA techniques are opening up a major new field of research in molecular evolution. Although these studies are an extension of previous phylogenetic analyses of protein sequences, total nuclear DNA and repetitive DNA sequences, they are showing much more precisely how genes have evolved and how new genetic functions are generated in evolution. This review will be concerned primarily with recent studies on the evolution of single copy animal genes and gene clusters, made possible by recombinant DNA methods. Earlier studies on the molecular evolution of proteins and repetitive DNA sequences have been fully reviewed elsewhere (Hood *et al.*, 1975; Tartof, 1975; Wilson *et al.*, 1977; Kedes, 1979; Long and Dawid, 1980).

II Experimental and theoretical approaches

A Detection and isolation of related genes

Most single copy genes examined to date have been detected using a cloned complementary DNA (cDNA) made from the required messenger RNA (see J. G. Williams, this series, Vol. 1). Labelled, cloned cDNA can be used to detect the corresponding homologous gene in restriction endonuclease digests of total genomic DNA, by agarose gel electrophoresis of DNA fragments followed by Southern blotting and filter hybridization (Southern, 1975, 1980). Similarly, cloned cDNA can be used to screen a recombinant λ bacteriophage-genomic DNA library, and to identify a recombinant plaque containing the corresponding gene (see R. A. Flavell, this volume).

Cloned cDNAs or genes can also be used to detect related gene sequences both in Southern blot hybridizations and in recombinant phage library screens. Detection will depend on the stringency of hybridization and the degree of DNA sequence divergence between the probe and the gene. For example, human β-globin cDNA detects the closely related β- and δ-globin genes at high stringencies of hybridization, plus the less closely related fetal globin genes ($^G\gamma$ and $^A\gamma$) at low stringencies (Flavell *et al.*, 1978). The maximum DNA sequence divergence between probe and gene compatible with detection of the gene in Southern blot analyses is about 25–30% over the region of greatest sequence homology. Somewhat greater levels of mismatch may be tolerated in plaque screenings of recombinant phage libraries, since the relative concentration of the relevant gene sequence in a plaque is greater than in genomic DNA on a Southern blot filter.

Cloned cDNAs and cloned gene sequences can also be used to detect and isolate homologous genes in the DNA of related species. Success will depend upon the divergence time of the two species and on the rate of evolution of coding sequences, which in turn is likely to be correlated with the clock rate of polypeptide sequence divergence. Thus rabbit adult β-globin cDNA is capable of detecting the entire family of human β-related globin genes (the ϵ-, $^G\gamma$-, $^A\gamma$-, δ- and β-globin genes, Barrie *et al.*, 1981), and has been used to isolate the human ϵ-globin gene from a recombinant λ phage library (Proudfoot and Baralle, 1979). Similarly, the chicken preproinsulin gene was isolated using a rat insulin cDNA probe (Perler *et al.*, 1980) and reptilian δ-crystallin DNA has been detected with a chicken probe (Williams and Piatigorsky, 1979). The most extreme cases of successful heterologous hybridization have been with cloned DNAs coding for highly conserved proteins. For example, sea urchin actin genes have been analysed using a *Drosophila* actin DNA probe (Durica *et al.*, 1980).

B Estimating DNA sequence divergence

If two related cloned fragments of DNA are to be compared for sequence divergence, the areas of homology must be located. This can be achieved by detailed restriction endonuclease mapping, cross-hybridization, or most speedily by electron microscopy of heteroduplexes formed between the two sequences. The latter method also gives a useful pictorial display of conserved and divergent sequences around related genes and has been used to compare, for example, the mouse β_d^{maj}- and β_d^{min}-globin genes (Leder *et al.*, 1980) and two chicken δ-crystallin genes (Jones *et al.*, 1980).

Once homologous sequences have been located, DNA sequence divergence is most accurately determined by total sequence analysis. The recent development of rapid DNA sequencing methods (Sanger *et al.*, 1978; Maxam and Gilbert, 1980) has led to an explosion of primary data suitable for evolutionary studies (Grantham *et al.*, 1980a, b, 1981). This situation is analogous to the appearance of polypeptide sequences during the 1960s, and will necessitate the development of central data banks capable of storing and updating complete sequences, and melding new partial sequences as they appear. Such banks are currently being assembled by the US National Institutes of Health (Bethesda) in co-operation with the European Molecular Biology Organization (Heidelberg), and also by Dr M. O. Dayhoff (Washington DC) and Prof. R. Grantham (Lyon) (see *Nature* 289, 112 (1980)).

Once two related DNA sequences are available, DNA sequence divergence can be calculated simply by aligning the sequences and scoring the proportion of nucleotides which differ between the two sequences. This procedure is most reliable when the two sequences are known to have diverged only by base substitution, as is generally the case when comparing two homologous protein coding sequences. Problems arise when deletions or insertions have occurred, as is frequently found in the evolution of extragenic DNA and intervening sequences. Optimal sequence alignment is commonly performed by constructing a matrix of nucleotide-by-nucleotide comparisons between the two sequences (Konkel *et al.*, 1979; see Fig. 1).

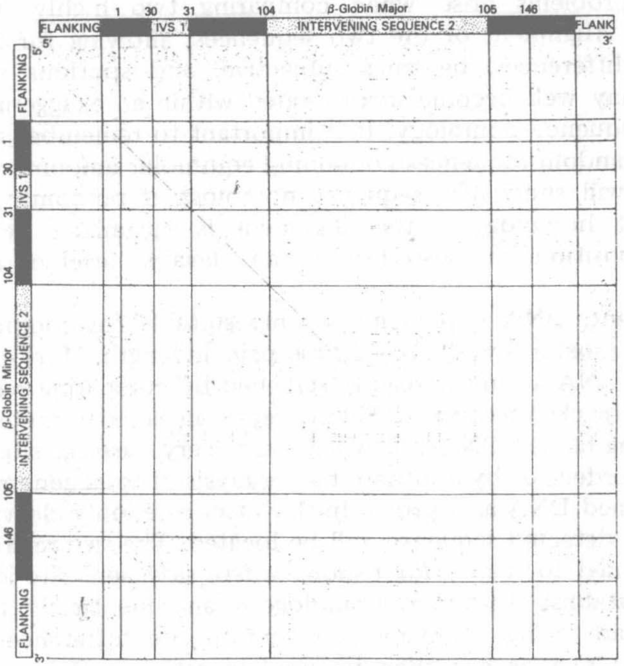

Figure 1 Dot matrix comparison of the DNA sequences of the mouse β_d^{maj}- and β_d^{min}-globin genes. The entire coding sequences, intervening sequences, and the immediate 5′ and 3′ flanking sequences are compared. Each dot represents the centre of a three-base identity between the two genes. Homology between these genes appears as a line at −45° across the grid. Discontinuities in the line in the second intervening sequence indicate gene divergence by microdeletion/insertion, as well as by base substitution (see Fig. 7). From Konkel *et al*., (1979); reprinted with permission of MIT press.

If the two sequences are identical, then a continuous line of identical bases appear on the diagonal, plus a scatter of spurious base identities over the entire matrix. Many of these unwanted identities can be removed by filtering out matches confined to an isolated base or pair of bases. If two sequences differ only by base substitutions, then the diagonal matrix line will show gaps corresponding to base differences. If a deletion or insertion has occurred, then the diagonal will show a discontinuity after which the line of homology will reappear parallel with the original line of identity. A measure of DNA sequence divergence can then be estimated by ignoring DNA tracts present in one sequence but not in the other. If several deletions/insertions have occurred, then a second independent estimate may be made of the frequency of these events per unit length of DNA. Other numerical methods of sequence alignment have been described by Van Ooyen *et al*. (1979).

Major problems arise when comparing two highly divergent sequences. Alignment of the two sequences, allowing for deletion/ insertion differences, becomes subjective, and spurious sequence matches may well become incorporated within an exaggerated estimate of sequence homology. It is important to remember that two mutually random sequences containing equimolar amounts of A, T, G and C will show 25% sequence homology if randomly aligned, and > 25% homology if the alignment is optimized. A skewed base composition can also lead to an elevated level of spurious homology.

At present, DNA sequencing is only suitable for comparing sequences at most a few thousand base pairs in length. More extensive regions of DNA are more readily screened by comparing restriction endonuclease cleavage maps. These maps can be constructed either by analysing cloned DNAs, in which case every cleavage site present will be recorded, or by Southern blot analysis of total genomic DNA using a cloned DNA as a probe. In the latter case, only cleavage sites nearest the detected sequence will be located. The two related maps can be aligned to search for regions where sufficient site identities exist to establish that two homologous and identically arranged sequences are being compared. Restriction site differences within homologous regions can then be used to estimate DNA sequence divergence.

Upholt (1977) gave the first description for converting restriction endonuclease cleavage map differences to sequence divergence. More refined methods both for this conversion and for correcting divergence estimates for multiple substitutions (see below) have been developed by Nei and Li (1979). Consider two homologous sequences, x and y. n_x cleavage sites will be examined in x, and n_y in y, of which n_{xy} will be common to both. The proportion, S, of sites present in one map, which are also present in the other map, is given by:

$$S = \frac{2n_{xy}}{n_x + n_y}$$

If restriction endonucleases with hexanucleotide recognition sequences have been used, S can be related to the DNA sequence divergence, λ, by:

$$\lambda = 1 - S^{1/6}$$

(As pointed out by Nei and Li (1979), ambiguities in the original definition of S by Upholt (1977) have resulted in several erroneous λ values appearing in the literature.) This analysis can be extended to endonucleases that recognize sequences other than hexanucleotides, and to nearest-neighbour maps, derived by Southern blot analysis