*Journal of Medical Genetics* (1972). **9**, 254.

# Gene Duplication, Mutation Load, and Mammalian Genetic Regulatory Systems*

SUSUMU OHNO

*Department of Biology, City of Hope Medical Center, Duarte, California, USA*

In recent years, the complete amino-acid sequences of increasing numbers of polypeptide chains became known. We have been afforded opportunities to look at the direct products of genes and this enabled us to deduce the evolutional history of individual gene loci. The extremely conservative nature of natural selection became immediately apparent. The observation that histone IV (110 amino-acid residues) of cattle and garden peas differ from each other only by two substitutions is most revealing (DeLange and Smith, 1971). It appears that an entire molecule of histone IV represents a functionally critical active site, so that almost any mutational amino-acid substitution makes a mutant protein delinquent in the performance of its assigned function which is to attach itself side-by-side to a DNA strand. Accordingly, natural selection has eliminated almost all the bearers of mutations affecting this gene locus since the creation of eukaryotes. Table I supports the view that the active site sequence of any functional polypeptide chain tended to remain invariant, as it appears that the smaller the proportion of active sites in a peptide chain, the faster the rate of its evolution. Fibrinopeptides A (15 to 19 amino-acid residues) and B (14 to 21 residues) are the fastest evolving of all the known peptide chains (Blomback, Blomback, and Grondahl, 1965; Dayhoff, 1969). During blood clot formation, fibrinopeptides A and B are discarded from the inert fibrinogen molecule by the trypsin-like action of thrombin. Their role being passive, most amino-acid substitutions would be harmless as only the carboxyl terminal arginine and a few other sites are functionally critical for their assigned function. There is indeed much evidence to support the view that most evolutional amino-acid substitutions represent functionally neutral, and therefore, trivial mutations (Kimura and Ohta, 1971).

So long as a particular function is assigned to a single gene locus in the genome, natural selection never permits character-changing mutations to affect that locus, for such mutations are deleterious to individuals which bear them. It follows then that natural selection is not a great advocator of changes as the Darwinian concept of evolution had us believe, but rather an extremely conservative force which tends to maintain the *status quo* at each locus. *Conditio sine qua non* of evolution had to be the escape from the relentless pressure exerted by natural selection. Only a redundant copy of an original gene created by the mechanism of gene duplication escapes from the stranglehold by natural selection, and while being ignored by natural selection, it is free to accumulate formerly forbidden mutations which change the active site. As a result, it may emerge as a new gene locus with a previously nonexistent function. There is little doubt that the creation of many new gene loci via gene duplication through polyploidy as well as unequal crossing-over contributed greatly to evolution from fish to mammals (Ohno, 1970).

Yet this mechanism too has its limitations. Since 3 of the 64 codons are chain terminating *nonsense* codons, there is a 1 in 24 chance that the first mutation sustained by a redundant copy would be chain-terminating. If long ignored by natural selection, a redundant copy would surely become a worthless degenerate DNA base sequence. In order not to degenerate, it has to acquire a new and useful function before long and begin contributing to the well-being of an organism. Only then, can it again be placed under surveillance by natural selection. It is natural selection that prevents further accumulation of randomly sustained mutations by this new

Received 24 April 1972.

* This work was supported in part by a grant (CA 05138) from the National Cancer Institute, US Public Health Service.

gene locus which would certainly lead to degeneracy. By doing so, natural selection in this role becomes an active contributor to evolutional changes. Nevertheless, a group of genes which shared a common ancestral gene, as a rule, show only a limited degree of functional divergence; ie, myoglobin and haemoglobin genes (Ingram, 1963). It becomes quite clear that evolution cannot be understood by knowing only the functional diversification of structural genes by gene duplication.

Indeed, we realize that major steps in vertebrate evolution were more often accomplished by changes in regulation of already existing structural genes rather than by the acquisition of new structural genes. The first major anatomical improvement that occurred to early vertebrates was the development of jaws. The earliest known vertebrates of more than 300 million years ago were jawless fish belonging to the class *Agnatha*. These fish-like creatures had a mouth which was merely an opening that led to the digestive tract with as many as 10 pairs of gills opening into it. Each gill was supported by an arch formed by a series of bones arranged in the fashion of a V turned on its side. When such a V of the third gill arch was strengthened and supplied with teeth and hinged at the point of the V, it became the jaw. Such a change required no new structural genes, but a change in direction of ontogenic development which is no doubt under the control of genetic regulatory systems. During mammalian evolution, extensive modification of digits occurred. Yet, the same set of structural genes are involved in the formation of man's hand, the horse's middle toe, and the whale's flipper. Thus, we come to realize that there is more truth than meets the eye to the time honoured statement 'ontogeny recapitulates phylogeny'. In order to know both processes, some understanding of the nature of genetic regulatory systems which operate in eukaryotes, in general, and in higher animals in particular is essential.

It has often been stated that compared with the *lac*-operon system of *E. coli* (Jacob and Monod, 1961) and other regulatory systems of prokaryotes, genetic regulatory systems in mammals must be enormously complicated, defying meaningful analysis at present. Such a modest statement has not necessarily been conducive to making headway in this direction, for while confessing ignorance, regulatory function of some sort has casually been assigned to undefined nonhistone proteins of the nucleus as well as to nuclear RNAs which do not belong to the known categories: *messenger*, *transfer*, and *ribosomal* RNAs. Similarly, a number of structural mutations in which amino-acid substitutions could not be demonstrated have been claimed to be regulatory in nature. It appears that we have been worse off for not having thought out the conceptual framework on the nature of mammalian regulatory systems. Any regulatory system whether of prokaryotes or of mammals must abide by certain rules and function within a framework imposed by certain restrictions. In the present paper, I shall try to point out the more obvious of these rules and restrictions.

Needless to say, the structural gene products such as enzymes are to a certain extent self-regulatory in their function because the concentrations of substrates and products as well as other conditions determine their efficiency. However, we are concerned here only with the genetic regulatory systems which direct ontogenic development of a fertilized egg. Jacob and Monod (1963) defined differentiation as '... two cells are differentiated with respect from one another if, while they harbour the same genome, the pattern of proteins which they synthesize is different'. One can think of no better definition.

## Mutation Load and Regulatory Systems

It is generally thought that the mammalian genome must harbour an enormous number of structural gene loci and that, in order to regulate such a large number of structural loci, mammals must possess infinitely complicated regulatory systems. The point is that an organism cannot afford to maintain an inordinately large number of gene loci. As soon as the copying mechanism of DNA replication based on the inherent complementality between pairs of bases (adenine and thymine, guanine and cytosine) becomes so perfect that there is no more room for error, the evolutional process will cease to exist. Errors (base substitutions, deletions, etc) do occur spontaneously and randomly and these errors which affect individual gene loci are recognized as mutations. So long as a given gene locus contributes to the well-being of an organism, mutations affecting it can be deleterious as well as neutral or advantageous to individuals which bear them. Natural selection eliminates the bearers of deleterious mutations as they are unfit. For some gene loci such as the histone IV locus, almost all mutations are deleterious, while for others, a considerable fraction of the mutations would be harmless (Table I). Nevertheless, there is fairly good agreement that for man and other mammals, the mean spontaneous deleterious mutation rate per locus per organism generation is of the order of $10^{-5}$. It follows then that the more gene

TABLE I

INFLUENCE OF 'ACTIVE SITE' ON MOLECULAR EVOLUTION

| | Amino-acid Site/Year | Peptide Chain/Year |
|---|---|---|
| Histone IV | $0{\cdot}0006 \times 10^{-8}$ | $0{\cdot}006 \times 10^{-8}$ |
| Cytochrome C | $0{\cdot}03 \times 10^{-8}$ | $3{\cdot}12 \times 10^{-8}$ |
| Haemoglobins $\alpha$ and $\beta$ | $0{\cdot}12 \times 10^{-8}$ | $17{\cdot}2 \times 10^{-8}$ |
| *Immunoglobulin κ-chain* | | |
| Variable region | $0{\cdot}34 \times 10^{-8}$ | $37{\cdot}4 \times 10^{-8}$ |
| Constant region | $0{\cdot}40 \times 10^{-8}$ | $44{\cdot}0 \times 10^{-8}$ |
| Fibrinopeptides A and B | $0{\cdot}90 \times 10^{-8}$ | $16{\cdot}2 \times 10^{-8}$ |

The evolutionally permissible mutation rate of each protein is expressed as the rate of amino-acid substitution, deletion, and insertion per year per amino-acid site as well as per peptide chain. Modelled after Dayhoff (1969).

loci an organism possesses in its genome, the higher the overall deleterious mutation rate. This is what Haldane (1957) explained as the cost of natural selection and Muller (1967) called the mutation load. Obviously, the mutation load imposes a finite upper limit to the number of gene loci an organism can afford to have.

The mammalian genome (haploid set of chromosomes) contains roughly $3 \times 10^{-9}$ mg of DNA which conveniently corresponds to $3 \times 10^{9}$ base pairs. Taking the average gene size to be $10^{3}$ base pairs which can specify a 330 amino-acid residue long polypeptide chain, there is room for $3 \times 10^{6}$ gene loci. The overall deleterious mutation rate per organism generation for $3 \times 10^{6}$ gene loci is 30, and recessive deleterious mutations accumulate in the genome generation after generation. Because of this, Kimura (1968) estimated that if the human genome contains that many gene loci, at this time in evolutional history, in order to ensure that a few of them would survive, each mated pair would have to produce $10^{78}$ zygotes. The realistic number of functionally significant gene loci in the mammalian genome has been estimated to be between 4 and $5 \times 10^{4}$ (Muller, 1967; Crow and Kimura, 1970). These many gene loci account for no more than 2% of the total genomic DNA. Although *ribosomal* and *transfer* RNA genes which exist in multiple copies have to be added to the above number, it indeed appears that the bulk of our genomic DNA has no definable function in that these segments are either never transcribed or if transcribed are not translated to meaningful amino-acid sequences. Accordingly, these segments are undergoing very rapid evolutional changes by unrestricted accumulation of randomly sustained mutations (Walker, 1968). One of the main reasons that caused the accumulation of so many nonfunctional base sequences in the mammalian genome is, I believe, a series of gene duplications that occurred in the past (Ohno, 1970). In order to emerge as a new gene locus with a previously nonexistent function, a redundant copy of a functional gene created by gene duplication had to be ignored by natural selection so that it could accumulate formerly forbidden mutations as mentioned in the introduction. The more likely fate of an ignored copy, however, is degeneration rather than a new functional locus. My estimate is that for every copy which emerged triumphant as a new gene locus, 10 or so copies joined the ranks of nonfunctional DNA. Thus, in order to double the number of gene loci, the genome size had to increase 10 times. This was the price which had to be paid for acquiring increasingly complex body organization. A considerable portion of nonfunctional DNA exists as long tandem repeats of a short sequence known as satellite DNA (Southern, 1970) occupying the heterochromatic region adjacent to the centromere of each chromosome (Jones, 1970). The origin of this repetitious DNA is not clear. In the euchromatic region, long stretches of nonfunctional DNA appear to space apart functional genes. This is reflected in the calculation that in mammals one genetic cross-over unit corresponds to $10^{6}$ base pairs. These spacer DNA sequences probably represent extinct species of structural genes in that they are redundant copies which failed to acquire new functions. Furthermore, it appears that parts of nonfunctional DNA base sequence adjacent to a structural gene are transcribed together with a structural gene to *messenger* RNA. Thus, *messenger* RNA of a given peptide chain tends to be two or three times longer than expected. *Messenger* RNAs apparently contain a long stretch of monotonous polyadenylic acid sequence, and the recent discovery of longer than usual mutant human haemoglobin alpha- as well as beta-chain (Flatz *et al*, 1971; Milner, Clegg, and Weatherall, 1971) indicates that the $3'$ end of haemoglobin *messenger* RNA normally contains a stretch of normally untranslatable sequence which is not poly A.

The above considerations on the number and spacing of functional genes in the mammalian genome are very pertinent to our discussion on the nature of mammalian genetic regulatory systems as we shall shortly see.

First of all, the mammalian genome does not appear to contain an exorbitantly large number of gene loci. Since the difference between *E. coli* and mammals with regard to the actual number of gene loci is likely to be 10-fold rather than a 1000-fold, mammalian regulatory systems need not be so complicated.

Secondly, not only structural genes, but also regulatory components, can sustain deleterious mutations. Thus, regulatory components contribute heavily to the overall mutation load. A regulatory locus ($i$) can mutate to either $i^s$ or $i^c$ and give the *noninducible* or *constitutive* mutant phenotype. In the case of a repressive locus, the former ($i^s$), in simplest terms, represents a mutational decrease by a repressor of its binding affinity to an inducer. All the structural gene products of a system which are normally *inducible* would not be made even in the presence of an inducer. The latter represents either a mutational loss of the binding affinity to the *operator* ($i^c$) or the loss by deletion of $i$ gene ($i^-$). Mutational base changes in the *operator* region affect its binding affinity to the wild-type regulator and give either the $o^s$ (*operator noninducible*) or $o^c$ (*operator constitutive*) phenotype. As a single regulatory locus controls a number of structural genes, individual regulatory mutations should be even more deleterious than individual structural mutations.

During evolution from unicellular prokaryotes to multicellular eukaryotes of increasing complexity, the necessary increase in number of regulatory systems had to be compensated by simplifying each regulatory system. If the increase in number of systems was accompanied by the increase in components of each system, the overall mutation load would have become unbearable. I contend, therefore, that the direction of natural selection has been to reduce the number of components in each regulatory system as multicellular organisms attained higher degrees of complexity (Ohno, 1971).

Thirdly, mammals may come to rely more heavily on translational controls rather than transcriptional controls.

In transcriptional control of prokaryotes, a regulator binds to the *operator* region on DNA, as the *operator* is either a part of or situated immediately adjacent to the *promotor* region (Smith and Sadler, 1971), this binding either prevents (repressive control) or enhances (activating control) the transcription of that regulated structural gene by RNA polymerase. Thus, the regulation is primarily transcriptional.

In mammals whose genome is so loaded with nonfunctional DNA base sequences, however, the precise transcriptional control may have become very difficult or nearly impossible. Indeed, the transcription in mammals appears to be rather indiscriminate as already indicated. The original observation (Harris and Watts, 1962) that up to 90% of the freshly transcribed RNAs in the mammalian cell nucleus (heterogeneous nuclear RNA) are instantly degraded either *in situ* or in the cytoplasma (Aronson, 1972) has been repeatedly confirmed (Harris, 1970). To be sure, there would still be a general sort of transcriptional control. For example, a steroid hormone induced hypertrophy of target cells appears to be caused in part by the selective activation of RNA polymerase I which results in increased production of *ribosomal* RNA, therefore, ribosomes (Liao *et al*, 1965; Blatti *et al*, 1970).

Nevertheless, precise transcriptional control as such may be the exception rather than the rule. Very recently, it became possible to make a DNA copy of a particular species of *messenger* RNA by the use of viral reverse transcriptase (Kacian *et al*, 1972; Verma *et al*, 1972). If it is shown in the near future that heterogeneous nuclear RNAs isolated from fibroblasts contain a sequence complementary to haemoglobin DNA, it will become certain that there is not much transcriptional control *sensus stricto* in mammals. A principle difference in regulatory systems that exists between prokaryotes and mammals might be that mammals principally exercise translational control on already made *messenger* RNAs.

## Essential Components of a Self-contained Regulatory System

There is little doubt that the construction of a mammalian body requires a rather large number of regulatory systems and these regulatory systems sort themselves out to sets of hierarchial arrangements. Furthermore, several independent regulatory systems may simultaneously operate in the same somatic cell type in an interlocked manner in that the same structural locus may be placed under the control of two different regulatory systems. For example, tyrosine aminotransferase in the rat liver can be induced either by hydrocortisone or by insulin (Lee and Kenney, 1971). Yet, in order to analyse a regulatory system, it has to be recognized as a self-contained entity distinguishable from others which are interlocked with it.

A regulatory system is definable as one which is under the control of a single regulatory locus, and it is likely to include a number of structural genes which are regulated. How can one regulatory gene control more than one structural gene ? In order to be regulated, all the structural genes or their *messenger* RNAs included in the system must share in common a stretch of identical or nearly identical DNA or RNA base sequence, for it is this base sequence to which a regulatory gene product has specific binding affinity. Such a base sequence which constitutes a part of each regulated structural gene is defined as the *operator* region. The *operator* region need not be and should not be long. The longer the region, the higher the possibility of sustaining deleterious mutations. Furthermore, a stretch made of 20 consecutive bases or base pairs can be sufficiently specific to function as the binding site for a particular regulatory gene product, for a given sequence represents one of $4^{20}$ possible sequences (Smith and Sadler, 1971).

If a binding between a regulatory gene product and the *operator* region results in preventing either transcription of that structural gene or translation of its *messenger* RNA, that product is functioning as a repressor and the regulation is repressive. If, on the other hand, this binding results in enhancing either transcription or translation, the activating control is in operation. A difference between the repressive and activating controls is not as great as it may appear at first glance.

Contrary to various claims made on acidic proteins and RNAs of mammalian cell nuclei, the mere indication or even the fact of having binding affinity to a particular group of structural genes does not suffice as a requirement to be a regulator of a self-contained regulatory system. For example, the presence in cell nuclei of *activator* RNA has been postulated. This smallish piece of single stranded RNA is said to complex with a particular DNA base sequence (*operator* ?), and by so doing removes histones from its vicinity and permits transcription of certain structural genes (Bekhor, Bonner, and Dahmus, 1969). Even if *activator* RNAs exist, the gene which specifies such RNA disqualifies itself as a regulator, since whether a regulatory system is to be switched on or off is dependent upon the presence or absence of *activator* RNA. A true regulatory gene of such a system must be the one which controls the gene which specifies *activator* RNA (Britten and Davidson, 1969).

A chromatid of an average mammalian chromosome 5$\mu$ or so in length should contain a DNA strand 4–5 cm long. A gene is but a particular portion of such a strand equipped with a short *promotor* region on one side and a transcription terminating signal on the other. RNA polymerase recognizes the *promotor* base sequence and initiates transcription. Although only two species of RNA polymerases, one for the transcription of *ribosomal* RNA genes in the nucleolus and the other for transcription of presumably all other genes, have so far been recognized in mammals (Roeder and Rutter, 1969); for the sake of argument, let us assume that the mammalian genome contains hundreds of gene loci for different species of RNA polymerases each having an affinity to a specific *promotor* base sequence. It follows then that each species of RNA polymerase transcribes only a particular group of structural genes. It may appear that each RNA polymerase locus is a master regulatory locus of a specific regulatory system. In fact, however, these polymerase loci are merely regulated structural loci as were the *activator* RNA loci.

From the above considerations, one comes to the conclusion that to be a self-contained regulatory system, it has to contain a regulatory locus which is itself not subjected to regulation in that system. It follows that a regulatory gene product by definition should be made constitutively (all the time). To put it another way, until a constitutively produced regulatory gene product is identified, that regulatory system has not revealed itself as a whole. The reason that neither *activator* RNA nor sequence specific RNA polymerase can be considered as having a primary regulatory function is that they have binding affinity only towards a specific base sequence (*operator* and *promotor*). Therefore, while they may have the power to switch on certain groups of structural genes, in order to do so, they themselves have to be switched on by some other means.

In order to be a master of a regulatory system, a regulatory gene product should have the capacity to modify its binding affinity to the *operator* region of structural genes it controls. Jacob and Monod (1961) reasoned that this modification occurs because a regulatory gene product also has binding affinity to a particular small molecule (inducer). If binding with an inducer reduces the regulator's affinity towards the *operator*, that regulator functions as a repressor. If, on the other hand, the inducer-bound regulator shows stronger affinity than unbound regulator towards the *operator*, it must function as an activator. Either way, a regulatory system becomes a self-contained unit in that it can assume either a switched on (*induced*) or switched off (*noninduced*) state depending upon the presence or absence of an inducer. Thus, by having such a regulatory system, the cell acquired the ability to adjust to changes in the outside environment which are represented as changing inducer concentrations.

Even in such complicated eukaryotes as mammals, the essential components of each self-contained regulatory system have to be (1) an inducer, (2) a regulator, and (3) the *operator* region of regulated structural genes or their *messenger* RNAs. As a regulator has to have binding affinity to both an

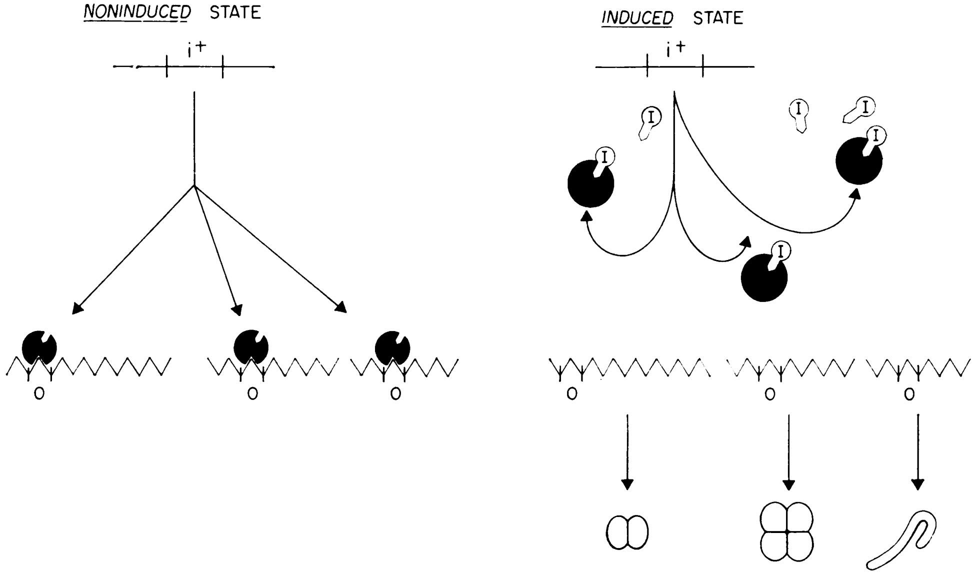

FIG. 1. Three essential components of any self-contained regulatory system. The system is expressed in its simplest form. (1) An inducer which is a small molecule such as a steroid hormone. (2) A regulator specified by *i* gene which has binding affinity to an inducer, on one hand, and to the *operator* base sequence of structural genes or their *messenger* RNAs it controls, on the other. (3) *Operator* (*o*) base sequence contained in the regulated structural genes or their *messenger* RNAs.

inducer and the *operator*, and since binding with an inducer should affect the regulator's affinity to the *operator*, a regulatory product has to be a protein rather than nucleic acid. These components are schematically depicted in Fig. 1. It is granted that there can be many modifications of the above scheme. Nevertheless, any regulatory system missing any of the essential components mentioned above is not self-contained, and only never ending circular arguments can sustain its purported existence.

## Simplification in Mammalian Regulatory Systems

Regulatory systems of prokaryotes utilize substrates or derivatives of substrates as inducers. For example, an inducer of the *lactose*-operon system is allolactose, a natural metabolite of lactose, or its artificial analogue, IPTG (isopropyl thiogalactoside). Accordingly, organisms pay no cost of natural selection for their inducers. Mammalian regulatory systems, on the other hand, utilize peptide hormones, steroid hormones, catecholamines, etc, as inducers. Thus, for synthesis of each inducer, a number of structural genes had to be set aside, and they constitute a part of the mutation load. This is yet another reason for simplification of each mammalian regulatory system. In the case of regulatory systems which utilize multitudes of peptide hormones as inducers, the simplicity was apparently achieved by introducing a common internal inducer as an intermediary. It appears that target cells of different peptide hormones differ from each other only by having a specific membrane-bound receptor for a particular peptide hormone. Nonetheless, the binding between a hormone and its receptor results in activating adenyl cyclase which cyclicizes adenosine monophosphate (AMP) by introducing a 3′ linkage, and the resulting cyclic AMP functions as an internal inducer (Robinson, Butcher, and Sutherland, 1968). Although cyclic AMP may or may not be the only internal inducer, the point is that economy was achieved by having a regulatory system which responds to a common inducer, instead of having separate regulatory systems for individual peptide hormones.

Steroid hormones, on the other hand, are incorporated into the cell and even into the nucleus. It appears that there is a separate regulatory system set up for each steroid hormone. As each system has to have a regulatory locus, the economy drive can only affect the *operator* region. As already mentioned, each *operator* region of a structural gene contributes to the mutation load. Even in the *lac*-operon system of *E. coli*, the fact that one *operator* region is shared by all 3 regulated structural genes rather than each having its own *operator* suggests that, from the very beginning of evolution, natural selection favoured the reduction in number of *operator* regions to reduce the mutation load of each regulatory system.

In the case of transcriptionally regulated structural genes, the *operator* region necessarily occupies a position adjacent to the *promotor* region where transcription by RNA polymerase is initiated. In the case of transcriptional control, positional requirement of the *operator* region disappears. Binding between a repressor and any part of *messenger* RNA may prevent its translation thereby enhancing its rate of degradation. Since a repressor molecule is not required to read the RNA base sequence as a series of triplet codons, *messenger* RNAs for polypeptide chains of very different function, and, therefore, different amino-acid sequences, may share in common a stretch of base sequence which is identical or nearly identical, thus coming under the control of a single repressor as illustrated in Fig. 2. In this way, the *operator* region as an independent entity has effectively been eliminated from the genetic regulatory system. By switching primarily to translational regulation, mammals may indeed have succeeded in reducing the mutation load of each regulatory system.

The same *operator* base sequence can be shared by functionally divergent *messenger* RNA

——GUUCGAAUUCCGCAGCCUGC——

(1) VAL·ARG·ILE·PRO·GLN·PRO
(2) PHE·GLU·PHE·ARG·SER·LEU
(3) SER·ASN·SER·ALA·ALA·CYS

FIG. 2. This shows that *messenger* RNAs which are translated to peptide chains of very different amino-acid sequences can contain a short stretch (20 bases or so long) of identical base sequence. Such a stretch can serve as an *operator*. Thus, *messenger* RNAs for very dissimilar enzymes such as alcohol dehydrogenase and $\beta$-glucuronidase can come under the control of a single regulatory protein.

Testosterone elicits two responses from its target cells; the induction of specific enzymes and hypertrophy. While hypertrophic or hyperplastic response is universal, different sets of enzymes are induced in different target organs. We studied the regulatory system of the mouse which responds to testosterone in kidney proximal tubule cells where 100-fold induction of alcohol dehydrogenase (EC1.1.1.1) and $\beta$-glucuronidase (EC3.2.1.31) is elicited. The actual inducers are intracellular metabolites of testosterone; $5\alpha$-dihydrotestosterone and $5\alpha$-androstane-$3\alpha$-$17\beta$-diol. The supernatant (cytosol) fraction of target cells contains a specific receptor protein to the former, and when the complex between a receptor and $5\alpha$-dihydrotestosterone is formed, the complex moves into the nucleus (Bruchovsky and Wilson, 1968; Fang, Anderson, and Liao, 1969). The latter in low concentration, on the other hand, has no binding affinity to a receptor protein in the cytosol fraction (Baulieu *et al*, 1971; Liao *et al*, 1971). Instead, in the microsomal fraction where *messenger* RNAs are, and, therefore, where a translational repressor should be, there exists another class of protein which has equal binding affinity to both $5\alpha$-androstane-diols ($3\alpha$ as well as $3\beta$) and $5\alpha$-dihydrotestosterone (Baulieu *et al*, 1971).

We found that, mg for mg, $5\alpha$-androstane-$3\alpha$-$17\beta$-diol is a more potent inducer of alcohol dehydrogenase and $\beta$-glucuronidase than $5\alpha$-dihydrotestosterone (Ohno, Dofuku, and Tettenborn, 1971). This suggests that response of target cells to testosterone is mediated by a translational regulator. Indeed, the induction of specific enzymes by steroid hormones in general appears to be due to the removal of translational rather than transcriptional block (Tomkins *et al*, 1969).

Has the independent *operator* region really been eliminated from the alcohol dehydrogenase and $\beta$-glucuronidase loci? The mouse genome contains only a single structural locus for $\beta$-glucuronidase located on the linkage group XVII chromosome (Paigen, 1963). This enzyme in other organs such as liver and spleen is made constitutively while in kidney proximal tubule cells, it is made inducibly only in the presence of testosterone. We have recovered an apparent amino-acid substituting mutation of this locus which simultaneously gave an $o^s$ (*operator noninducible*) character with regard to its inducibility by testosterone in kidney proximal tubule cells. Constitutive levels of this mutant $o^sG^s$ (*operator noninducible* slow electrophoretic variant) enzyme in various nontarget cells remained the same as those of the wild-type $o^+$G enzyme (Dofuku, Tettenborn, and Ohno, 1971a and b). Thus, the view that a translatable part of *messenger* RNA serves as the *operator* region finds a measure of support.

The logical extension of the argument that each mammalian regulatory system must be simple is that the developmental fate of an entire organ or even a series of organs might be placed under the control of a single regulatory locus. Our findings on the X-linked testicular feminization (*Tfm*) mutation of the mouse (Lyon and Hawkes, 1970) indeed indicates that this can be the case (Ohno, Stenius, and Christian, 1970). In a series of embryological experiments on rabbits and rats, Alfred Jost (1961) has shown that the expression of the male or female phenotype strictly depends on the presence or absence of testosterone. This is because the persistence and differentiation of embryonic Müllerian ducts to Fallopian tubule and uterus of the female are a constitutive process which does not require an inducer, while the mesonephric (Wolffian) duct can persist and differentiate to epididymis and seminal vesicals of the male only in the presence of an inducer (testosterone). In the absence of testosterone, the urogenital sinus too would differentiate towards the female direction and form vagina and vulva instead of penis and prostates. Thus, it can be said that maleness and femaleness represent the *induced* state and *noninduced* state of one and the same regulatory system. If that were so, an $i^s$ (*regulator noninducible*) mutation of this regulatory locus should give the female phenotype to a mutant even in the presence of testosterone. Indeed, the X-linked *Tfm* mutation of the mouse which genetically behaves as a point mutation of a single locus satisfies all the requirements to be an $i^s$ mutation. Affected $X^{Tfm}Y$ chromosomal males are equipped with testes which produce a normal amount of testosterone at least until the neonatal stage. Yet, derivatives of the mesonephric ducts are completely absent, and their external phenotype is female. Neither induction of alcohol dehydrogenase and β-glucuronidase nor hypertrophy occurs in $X^{Tfm}Y$ kidney proximal tubule cells even after the administration of as much as 20 mg of testosterone, 5α-dihydrotestosterone and 5α-androstane-3α-17β-diol (Ohno and Lyon, 1970; Dofuku *et al*, 1971a).

Analogous mutations of other regulatory loci which control the developmental fate of other organs have not been found in the mouse and other mammalian species. There is a very simple explanation for this apparent absence. Such mutations of other regulatory loci would almost surely cause embryonic death, thus they would be classified as embryonic lethals and escape detection as regulatory mutations. A sexual phenotype is a luxury without which an individual can survive, and this enabled us to study the *Tfm* mutation.

## Certain Conditions to be Met by a Regulatory Protein

We have already discussed certain characteristics which a regulatory gene product has to have. It should have binding affinity to an inducer, on one hand, and to the *operator* base sequence on the other, and binding with an inducer should change its binding affinity to the *operator*. Thus, a regulator is almost certainly a protein rather than RNA.

In target cells of each steroid hormone a class of proteins having a high binding affinity to that steroid exists. In a strict sense, only one of them has to be a regulatory protein. Fortunately, there are a few rigid qualifications which a regulatory gene product has to fulfil and defining them will help us to single out a regulatory protein of a given system from a number of candidates. Since what applies to a repressive regulatory gene product applies equally well to an activating regulator, we will discuss these qualifications only for the repressor.

First of all, a regulatory gene product by definition should be made constitutively in the cell in which that regulatory system is operating. A protein which is inducible cannot be the product. Secondly, a regulatory protein should exist in an exceedingly low concentration. Although the *operator* region should show the highest binding affinity to its repressor, DNA and RNA in the cell are bound to contain other base sequences that also show varying degrees of binding affinity towards that repressor. Purified *lac*-repressor protein of *E. coli*, for example, shows a considerable binding affinity to poly dAT sequence (Lin and Riggs, 1970). Conversely, binding between an inducer and a repressor does not completely abolish the affinity between a repressor and the *operator*, it merely reduces the affinity by a few orders of magnitude (Riggs, Newby, and Bourgeois, 1970). Furthermore, the rate of inducible enzyme synthesis varies inversely with the first power of the repressor concentration (Sadler and Novick, 1965). Thus, the higher the concentration of a repressor, the more difficult it becomes to bring about the induced state.

It appears that in order to function as a specific regulator of certain structural genes, a regulatory gene product must exist in the cell in a very low concentration. Indeed, it has been estimated that the wild-type *E. coli* contains only 10 molecules of *lac*-repressor per cell (Müller-Hill, 1971). One concludes that any protein either in the nucleus or the cytoplasma that exists in rather large quantity (eg, several thousand molecules per cell), is not likely to be a regulatory protein.

Thirdly, the above relationship between the rate of inducible enzyme synthesis and the repressor concentration enables us to predict the degree of binding affinity (Km) a particular regulatory protein has to have towards its inducer. The wild-type *lac*-repressor protein of *E. coli* has a Km of $1{\cdot}36 \times 10^{-6}$ M towards an inducer (IPTG) and at this inducer concentration no induction occurs in permeaseless mutant *E. coli*. Half-maximal induction of three enzymes in the *lac*-operon occurs at $2 \times 10^{-4}$ M IPTG, and the maximal induction requires an inducer concentration greater than $10^{-3}$ M (Gilbert and Müller-Hill, 1966).

In mammals, 93 to 96% of the circulating testosterone in blood plasma is protein-bound, and, therefore, may not contribute to the inducer concentration. The free testosterone concentration of normal man is about $3{\cdot}6 \times 10^{-10}$ M and that of normal woman about $1{\cdot}7 \times 10^{-11}$ M (Hudson *et al*, 1970). The same relationship holds true for the mouse, although the respective concentrations are slightly above one third of those of man. Both marker enzymes in normal female kidney proximal tubule cells are in a noninduced state, while they are in 30% induced state in the normal male and the maximal induction requires a single injection to an adult mouse of either 1 mg of 5α-androstanediol or 3 mg of 5α-dihydrotestosterone which would boost the inducer concentration above $10^{-8}$ M (Ohno *et al*, 1971; Dofuku *et al*, 1971a). There is a remarkable parallel between the *lac*-operon system of permeaseless *E. coli* and a regulatory system controlled by the X-linked *Tfm* locus of the mouse in this respect. This comparison is fair because kidney proximal tubule cells do not appear to be equipped with a specific permease for testosterone (Bullock, Bardin, and Ohno, 1971). It follows that a regulatory protein specified by the *Tfm* locus should have a Km towards an inducer in the range of $10^{-11}$ M.

The point which I wish to make is that once a pertinent mutation or mutations are recovered, the identification of a regulatory gene product need not be a difficult task. A mutational reduction by a few orders of magnitude of the binding affinity towards an inducer, for example, a change from the wild-type Km of $10^{-11}$ M to a mutant Km of $10^{-9}$ M, suffices as the molecular basis of the $i^s$ (*regulator noninducible*) mutation.

## Regulatory Systems and Evolution by Gene Duplication

There is little doubt that evolutional increase in the number of regulatory systems was also accomplished by a successive series of gene duplications. There is at least one regulatory gene for each of the multitudes of steroid hormones the body produces. These regulatory genes must have shared a common ancestry sometime in the past, since the amino-acid sequence required for the specific binding with testosterone is bound to be more similar than different from that required for high affinity binding with, say, progesterone.

In the case of structural genes, too, a true functional diversification between a newly created gene and its immediately ancestral gene is accomplished when the two are placed under the control of different regulatory genes. This requires a new structural gene which is created by gene duplication to acquire an *operator* base sequence different from that possessed by its predecessor.

The observation that the fate of a series of organs derived from mesonephric (Wolffian) ducts is placed under the control of a single X-linked *Tfm* locus suggests to us that drastic evolutional changes of digits and other body parts may have been accomplished by a few changes in regulatory components. As we learn more about genetic regulatory systems of mammals and other higher organisms, we will begin to understand the true molecular basis of evolution as well as ontogeny.

### References

Aronson, A. I. (1972). Degradation products and a unique endonuclease in heterogeneous nuclear RNA in sea urchin embryos. *Nature New Biology*, **235**, 40–44.

Baulieu, E. E., Jung, I., Blondeau, J. P., and Robel, P. (1971). Androgen receptors in rat ventral prostate. In *Advances in the Biosciences*, ed. by G. Raspé, vol. 6. Pergamon Press, Oxford.

Bekhor, I., Bonner, J., and Dahmus, G. K. (1969). Hybridization of chromosomal RNA to native DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **62**, 271–277.

Blatti, S. P., Ingles, C. J., Lindell, T. J., Morris, P. W., Weaver, R. F., Weinberg, F., and Rutter, W. J. (1970). Structure and regulatory properties of eucaryotic RNA polymerase. *Cold Spring Harbor Symposium on Quantitative Biology*, **35**, 649–657.

Blomback, B., Blomback, M., and Grondahl, N. J. (1965). Studies on fibrinopeptides from animals. *Acta Chemica Scandinavica*, **19**, 1789–1791.

Britten, R. J. and Davidson, E. H. (1969). Gene regulation for higher cells: A theory. *Science*, **165**, 349–357.

Bruchovsky, N. and Wilson, J. D. (1968). The intranuclear binding of testosterone and 5α-androstan-17β-ol-3-one by rat prostate. *Journal of Biological Chemistry*, **243**, 5953–5960.

Bullock, L. P., Bardin, C. W., and Ohno, S. (1971). The androgen insensitive mouse: Absence of intranuclear androgen retention in the kidney. *Biochemical and Biophysical Research Communications*, **44**, 1537–1543.

Crow, F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Dayhoff, M. O. (ed.) (1969). *Atlas of Protein Sequence and Structure*, vol. 4. National Biomedical Research Foundation, Silver Springs, Maryland.

DeLange, R. J. and Smith, E. L. (1971). Histones: Structure and function. *Annual Review of Biochemistry*, **40**, 279–314.

Dofuku, R., Tettenborn, U., and Ohno, S. (1971a). Testosterone-regulon in the mouse kidney. *Nature New Biology*, **232**, 5–7.

Dofuku, R., Tettenborn, U., and Ohno, S. (1971b). Further characterization of the $o^s$ (*operator noninducible*) mutation of the mouse β-glucuronidase locus. *Nature New Biology*, **234**, 259–261.

Fang, S., Anderson, K. M., and Liao, S. (1969). Receptor proteins for androgens. On the role of specific proteins in selective retention of 17β-hydroxy-5α-androstan-3-one by rat ventral prostate *in vivo* and *in vitro*. *Journal of Biological Chemistry*, **244**, 6584–6596.

Flatz, G., Kinderlerer, J. L., Kilmartin, J. V., and Helmann, H. (1971). Haemoglobin Tak: A variant with additional residues at the end of the β-chains. *Lancet*, **1**, 732–733.

Gilbert, W. and Müller-Hill, B. (1966). Isolation of the lac repressor. *Proceedings of the National Academy of Sciences of the United States of America*, **56**, 1891–1898.

Haldane, J. B. S. (1957). The cost of natural selection. *Journal of Genetics*, **55**, 511–524.

Harris, H. (1970). *Nucleus and Cytoplasm*, 2nd ed. Clarendon Press, Oxford.

Harris, H. and Watts, J. W. (1962). The relationship between nuclear and cytoplasmic ribonucleic acid. *Proceedings of the Royal Society. Series B*, **156**, 109–121.

Hudson, B., Burger, H. G., de Kretser, D. M., Coghlan, J. P., and Taft, H. P. (1970). Testosterone plasma levels in normal and pathological conditions. In *The Human Testis*, pp. 423–437, ed. by E. Rosemberg and C. A. Paulsen. Plenum Press, New York.

Ingram, V. M. (1963). *The Hemoglobin in Genetics and Evolution*. Columbia University Press, New York.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanism in the synthesis of proteins. *Journal of Molecular Biology*, **3**, 318–356.

Jacob, F. and Monod, J. (1963). Genetic repression, allosteric inhibition, and cellular differentiation. In *Cytodifferentiation and Macromolecular Synthesis*, ed. by M. Locke. Academic Press, London.

Jones, K. W. (1970). Chromosomal and nuclear location of mouse satellite DNA in individual cells. *Nature*, **225**, 912–919.

Jost, A. (1961). The role of fetal hormones in prenatal development. *The Harvey Lectures*, Series 55. Academic Press, New York.

Kacian, D. L., Spiegelman, S., Bank, A., Terada, M., Metafora, S., Dow, L., and Marks, P. A. (1972). *In vitro* synthesis of DNA components of human genes for globins. *Nature New Biology*, **235**, 167–169.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Kimura, M. and Ohta, T. (1971). Protein polymorphism as a phase of molecular evolution. *Nature*, **229**, 467–469.

Lee, K.-L. and Kenney, F. T. (1971). Assessment of hormone action in cultured cells. In *Karolinska Symposium on Research Methods in Reproductive Endocrinology*, 3rd Symposium, ed. by E. Diczfalusy. Bogtrykkeriet Forum, Copenhagen.

Liao, S., Leininger, K. R., Sagher, D., and Barton, R. W. (1965). Rapid effect of testosterone on ribonucleic acid polymerase activity of rat ventral prostate. *Endocrinology*, **77**, 763–765.

Liao, S., Tymoczko, J. L., Liang, T., Anderson, K. M., and Fang, S. (1971). Androgen receptors: 17$\beta$-hydroxy-5$\alpha$-androstan-3-one and the translocation of a cytoplasmic protein to cell nuclei in prostate. In *Advances in the Biosciences*, ed. by G. Raspé, vol. 6. Pergamon Press, Oxford.

Lin, S. and Riggs, A. D. (1970). *Lac* repressor binding to DNA not containing the *lac* operator and to synthetic poly dAT. *Nature*, **228**, 1184–1186.

Lyon, M. F. and Hawkes, S. G. (1970). An X-linked gene for testicular feminization in the mouse. *Nature*, **227**, 1217–1219.

Milner, P. F., Clegg, J. B., and Weatherall, D. J. (1971). Haemoglobin-H disease due to a unique haemoglobin variant with an elongated $\alpha$-chain. *Lancet*, **1**, 729–732.

Muller, H. J. (1967). The gene material as the initiator and the organizing basis of life. In *Heritage from Mendel*, ed. by R. A. Brink. University of Wisconsin Press, Madison.

Müller-Hill, B. (1971). *Lac* repressor. *Angewandte Chemie* (international edition), **10**, 160–172.

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer-Verlag, Heidelberg.

Ohno, S. (1971). An argument for the simplicity of mammalian regulatory systems: Single gene determination of male and female phenotypes. *Nature*, **234**, 134–137.

Ohno, S., Dofuku, R., and Tettenborn, U. (1971). More about X-linked testicular feminization of the mouse as a noninducible ($i^s$) mutation of a regulatory locus: 5$\alpha$-androstan-3$\alpha$-17$\beta$-diol as the true inducer of kidney alcohol dehydrogenase and $\beta$-glucuronidase. *Clinical Genetics*, **2**, 1–13.

Ohno, S. and Lyon, M. F. (1970). X-linked testicular feminization in the mouse as a *noninducible* regulatory mutation of the Jacob-Monod type. *Clinical Genetics*, **1**, 121–127.

Ohno, S., Stenius, C., and Christian, L. (1970). Sex difference in alcohol metabolism: Androgenic steroid as an inducer of kidney alcohol dehydrogenase. *Clinical Genetics*, **1**, 35–44.

Paigen, K. (1964). The genetic control of enzyme realization during differentiation. In *Congenital Malformations*, Second International Conference, 1963. The International Medical Congress, Ltd, New York.

Riggs, A. D., Newby, R. F., and Bourgeois, S. (1970). *Lac* repressor-operator interaction. II. Effect of galactosides and other ligands. *Journal of Molecular Biology*, **51**, 303–314.

Robinson, G. A., Butcher, R. W., and Sutherland, E. W. (1968). Cyclic AMP. *Annual Review of Biochemistry*, **37**, 149–174.

Roeder, R. G. and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, **224**, 234–237.

Sadler, J. R. and Novick, A. (1965). The properties of repressor and the kinetics of its action. *Journal of Molecular Biology*, **12**, 305–327.

Smith, T. F. and Sadler, J. R. (1971). The nature of lactose operator constitutive mutations. *Journal of Molecular Biology*, **59**, 273–305.

Southern, E. M. (1970). Base sequence and evolution of guinea-pig $\alpha$-satellite DNA. *Nature*, **227**, 794–798.

Tomkins, G. M., Gelehrter, T. D., Granner, D., Martin, D., Samuels, H. H., and Thompson, E. B. (1969). Control of specific gene expression in higher organisms. *Science*, **166**, 1474–1480.

Verma, I. M., Temple, G. F., Fan, G., and Baltimore, D. (1972). *In vitro* synthesis of DNA complementary to rabbit reticulocyte 10S RNA. *Nature New Biology*, **235**, 163–167.

Walker, P. M. B. (1968). How different are the DNA's from related animals? *Nature*, **219**, 228–232.