

THE ORIGINS OF MOLECULAR EVOLUTION

I. INTRODUCTION

The advent of molecular biology has affected the study of evolution as much as any other area of biology. It accelerated the pace of evolutionary biology by providing new technology but also by forcing the development of better models of evolution. In so doing it revolutionized the study of evolution in many ways, two of the most important being:

1. Initially it was believed that the driving force of evolution was selection for advantageous mutations. Molecular evolution showed that sites varied in their rates of evolution, and the most rapidly evolving sites were assumed to be the most important for function of the gene product. The discovery of neutral mutations and better models of evolution showed that most of the time evolution is driven by mutation and random drift, and that the most rapidly evolving sites are the *least* important ones.
2. Systematists developed rigorous rules for making phylogenetic trees in order to put their science on a better footing and relate it to evolution. Molecular evolutionary biologists developed algorithms for applying these rules to the construction of trees from sequence data, and developed alternative rigorous rules. The methods and the trees are now being used not only to describe evolutionary history and inform systematics but also as tools to study evolutionary mechanisms.

II. LAYING THE GROUNDWORK

1858-9 Charles Darwin: evidence for evolution; natural selection as mechanism

1900-1916 Correns, DeVries, Boveri, Morgan, Muller, Sturtevant, Bridges, etc.: rediscovered Mendel's laws, developed chromosome theory, linkage, etc.

1908 **G. H. Hardy, Wilhelm Weinberg:** origin of population genetics (concept of gene (allele) and genotype frequencies; genotype frequencies under random mating model).

1920's, 1930's ... R. A. Fisher, Sewall Wright, J. B. S. Haldane: mathematical foundations of population and evolutionary genetics.

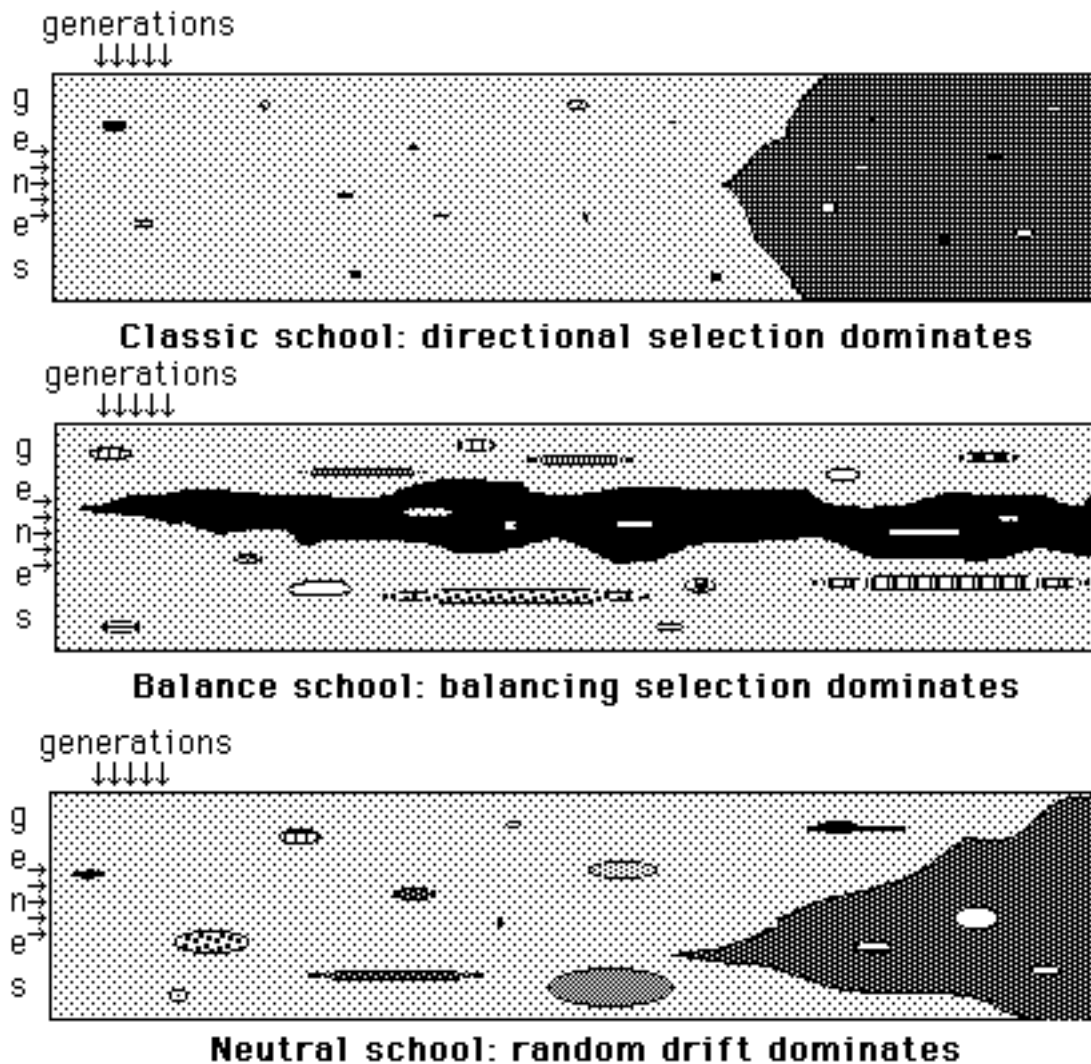
The phenotypes that were studied, and that geneticists tried to understand, were morphological (e.g. shape of wings of birds, size of mammals, number of bristles on a fruitfly, dark or light color of moths, number of leaves on a plant), aspects of fitness (e.g. fecundity, longevity, seed set), physiological traits (e.g. resistance to disease or toxic substances, metabolic rates), and occasionally biochemical (e.g. kinds of phenolic compounds in plants, coat pigments in mammals).

Explaining natural variation was handled differently by two schools of thought:

- (1) The Classic School (e.g. **H. J. Muller**) believed that for most genes most of the time, there is a single common allele (the "wild type") that confers maximum fitness in the present environment, and a few copies of a few other alleles that are detrimental and are being eliminated. Rarely, a new mutation is advantageous relative to the present wild type allele and replaces it by selection, but replacement is so rapid that polymorphism is transient.
- (2) The Balance School (e.g. **Theodosius Dobzhansky**) believed that many loci are polymorphic, with two or more alleles maintained at high frequency by selection for heterozygotes.

For both schools, the driving force of evolution was selection. If you wanted to explain an evolutionary change, you generally invoked selection for advantageous alleles or genotypes. There was little

discussion of the possibility of neutral mutations, i.e. mutations that do not affect the fitness of individuals. This is still true for adaptive changes at the morphological level. But now, at the molecular level, mutation and random drift are viewed as the driving forces for most changes.



III. THE MOLECULAR REVOLUTION

The molecular revolution in evolution began by switching phenotypes from morphology (and a bit of biochemistry) to sequences, initially protein sequences:

1952 Fred Sanger compared insulin sequences from cattle, sheep, horse, pig, and sperm whale. The results showed that some parts of the molecule were less variable than others, and Sanger concluded that the former were of less functional importance. (Sanger, F., 1952 *Adv. Protein Chem.* 7:1-67.)

1961 Vernon M. Ingram did the first molecular genealogy of a set of genes, the vertebrate globin gene family. From the partial amino acid sequences, Ingram proposed that myoglobin and the α , β , γ , and δ hemoglobins arose by successive gene duplications of a primitive globin gene (probably the first tree of genes based on sequences). On the basis of structure and function of the molecules, he argued that some mutations would change essential structural features and hence would not be "tolerated"

while others would, and still others would be selected; i.e. he considered the possibility that there would be neutral mutations as well as detrimental and advantageous mutations.

1962

Ernst Freese, a molecular biologist tried to explain why different species have different GC contents, while the GC content of different parts of a microbial genome have about the same GC content. He proposed that most mutations are selectively neutral because some DNA has no genetic function, some parts of a protein can be altered without affecting its function, and the code might be degenerate (this was not yet established).

Emile Zuckerkandl and Linus Pauling showed that the number of amino acid differences in a given protein increases ca. linearly with the divergence time of pairs of species, thus laying the empirical groundwork for parts of the neutral theory and the molecular clock hypothesis.

By the autumn of 1964, many of the major concepts of molecular evolution were in the air. The majority of codons had been identified and it was clear that the code was degenerate. Zuckerkandl and Pauling's proposal for rate constancy of amino acid substitution had been applied by others, and likewise attacked by classical evolutionary biologists such as George Gaylord Simpson and Ernst Mayr; the molecular clock debate had begun.

IV. The Revolution in Understanding the Roles of Mutation, Drift, and Selection

1966 Genetic diversity was measured at the molecular level by analyzing electrophoretic variants of enzymes in *Drosophila* (**Richard C. Lewontin and J. L. Hubby**, USA) and humans (**Henry Harris**, Great Britain). The large amount of diversity they found at the molecular level seemed to contradict the prediction of the Classic School. However, it also seemed to contradict the Balance School and the all-important role for selection in evolution. This was because the amount of genetic variation was so large that most organisms would be carrying a huge genetic load of detrimental mutations. These difficulties were partly solved later, but at the time they were joined to the accumulating body of molecular data to produce the neutral theory of evolution:

1968 Motoo Kimura showed mathematically that populations can maintain a high level of genetic diversity for neutral or nearly neutral alleles. From various experimental studies, and the fact that many amino acids have several synonymous codons, he proposed that as many as 40% of all mutations might be neutral or nearly neutral.

In a previous paper, Kimura and James Crow showed that selection for heterozygotes can only maintain heterozygosity at a small number of genes without increasing the genetic load to an impossibly high level.

In another paper, Kimura used the new measurements of the rate of amino acid substitution in vertebrates, and other data, to show that most mutations must be neutral or nearly neutral. He also noted that, if most mutations were neutral, then random drift must play a major role in determining the genetic structure of populations, i.e. their heterozygosity.

1969 Jack King & Tom Jukes marshalled an impressive number of additional arguments that most evolutionary changes in protein sequences are neutral.

It is important to emphasize that none of the proponents of the neutral theory denied the importance of selection for advantageous mutations for *adaptation*. The accumulation of neutral mutations was not used to explain how mammals in cold environments acquired warm fur coats or how flowers evolved mechanisms to attract the birds and bees. It was used to explain the majority of genetic diversity, and

the majority of evolutionary base substitutions and amino acid substitutions, that are seen at the molecular level. Even so, it was intensely controversial, and still is to some extent.

These papers completed the revolution in evolution, turning the way we looked at evolution upside down. The driving forces at the molecular level were now mutation and drift; selection had become a conservative force.

1977 1977 saw the development of rapid and simple DNA sequencing by Sanger et al., and by Maxam & Gilbert. Molecular evolution was off and running, and never looked back.

V. THE PHYLOGENETIC REVOLUTION

Late 1960s-1970s For several years, protein sequences had been used to reconstruct the evolutionary history of genes and organisms (phylogenetic analysis). At the same time, Willi Hennig and other systematists argued that systematics ought to be based on evolutionary relationships, especially on inclusive groups of organisms that share a common ancestor (clades). They also argued that clades should be identified by a rigorous procedure that we call parsimony analysis.

Computer algorithms for making trees from gene sequence data by parsimony and other methods (especially neighbor-joining and maximum likelihood) were developed by evolutionary biologists beginning in the late 1960s. These required good computer algorithms for aligning the sequences of a gene from different organisms, which were also developed. By the mid-1980s, phylogenetic trees were being used routinely in a variety of applications, e.g.:

1. Reconstructing evolutionary history of organisms with no fossil record and no good morphological features. This includes using the molecular clock to establish the dates of evolutionary events.
2. Reconstructing ancestral states, including gene and protein sequences.
3. Bringing rigor to comparative studies by identifying sister clades that differ in the trait of interest.
4. Detecting sexual reproduction and horizontal gene transfer.

Today phylogenetic analysis has become an absolute requirement for many kinds of research.

VI. THE GENOMICS AND BIOINFORMATICS REVOLUTION AND THE FUTURE

Genomics = the sequencing and analysis of large segments of genomes. Bioinformatics = the organization, storage, retrieval, and analysis of very large amounts of biological data (genomics data in this context).

Data are already enabling us to study the evolution of whole chromosomes and genomes, including their organization as well as their sequences. I predict that in the future we will use data from genomics and proteomics and associated methods and fields to identify rules that govern gene interactions (a new version of metabolic pathways) and the evolution of the interactions. We will uncover new rules governing the arrangement of genes on chromosomes and the constraints on the evolution of chromosome and genome evolution.