AIM

Association for Information and Image Management

1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910
301/587-8202

Centimeter

Inches

MANUFACTURED TO AIM STANDARDS
BY APPLIED IMAGE, INC.

ANL/BIM/PP--73690-Rev.2

ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439

GENOMIC DEFINITION OF SPECIES

by

Radomir Crkvenjakov & Radoje Drmanac
Biological and Medical Research Division

For submission to
Journal of Theoretical Biology

March 1993

RECEIVED
JUN 10 1994
OSTI

The submitted manuscript has been authored
by a contractor of the U. S. Government
under contract No. W-31-109-ENG-38.
Accordingly, the U. S. Government retains a
nonexclusive, royalty-free license to publish
or reproduce the published form of this
contribution, or allow others to do so, for
U. S. Government purposes.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

GENOMIC DEFINITION OF SPECIES

by

Radomir Crkvenjakov and Radoje Drmanac

INTRODUCTION

A genome is the sum total of the DNA sequences in the cells of an individual organism. The common usage that species possess genomes comes naturally to biochemists, who have shown that all protein and nucleic acid molecules are at the same time species- and individual-specific, with minor individual variations being superimposed on a consensus sequence that is constant for a species. By extension, this property is attributed to the common features of DNA in the chromosomes of members of a given species and is called (species) genome.

The definition of species based on chromosomes, genes, or genome common to its member organisms has been implied or mentioned in passing numerous times. Some population biologists think that members of species have similar "homeostatic genotypes," which are to a degree resistant to mutation or environmental change in the production of a basic phenotype (Mayr 1970;, Ehrlich and Raven 1969; Wiley 1981). Molecular biology textbook writers seem to imply the existence of species genomes (Alberts et al. 1989, Singer and Berg 1991). More

explicit statements can be found in the recent literature (Depew and Weber 1988; Colier 1988; Wesson 1991). We attempt here to outline the theoretical consequences of the view that genomes are simultaneously organism- and species-specific.

Another important concept is the genome program as a set of instructions for complex functions and development. From the operon concept in 1961, there has been a steady accumulation of evidence for the pattern of time and space control of gene expression, which is not only determined by the various sequence elements but also by their actual position and order in DNA. The alternative programs for the lytic and integrative life cycles of lambda phage, now completely understood on sequence level (Ptashne 1986), provide a paradigm for a single genome encoding a variety of alternative programs needed for the development of various forms in polytypic species (such as queens, drones, and workers in bees). The conservation of homeobox gene batteries in metazoan animals (Murtha et al. 1991) suggests the importance of DNA structures larger than genes in animal development. The surprising stabilization of one or few configurations in the simulation of behavior of genetic networks (Kauffman 1969) provides a conceptual answer for the question of how stable complex programs can be encoded in genomes. Now, it is even possible to propose simple models for genetic programs leading from

zygote to adult arthropod (Bodnar et al. 1989). Canalization (Waddington 1957) and developmental constraints (Alberch 1980) are concepts which describe the elasticity of development to perturbation and the restricted range of allowed outcomes of acceptable change. These systemic properties of developing organisms can be explained either on a nongenetic, cellular interaction level or can be directly or indirectly attributed to the species genome program encoded in the DNA sequences. In the definition proposed here, we have tried to clarify the relationship between the concepts of genome, DNA sequence, and genome program on a general level.

Three definitions of species have the most proponents today. They are the biological species concept (Mayr 1942) based on interbreeding potential, the ecological species concept based on niche occupancy, and the evolutionary (or phylogenetic) species concept based on lineages of organisms. A cogent critique of these concepts from the standpoint of genetics has been summarized recently (Templeton 1989). In the last decade, the philosophy of science has contributed to the ongoing debate about the correct definition of species. A single definition justifying the expectations of biologists and philosophers alike would have to satisfy three criteria (slightly modified from Kitcher [1987]):

UNIVERSALITY I. All organisms (evaluated over a full life cycle) must belong to exactly one species.

UNIVERSALITY II (Uniformity). A species category is characterized by principle(s) valid for all species taxa (individual species).

ULTIMACY. The principle(s) by which one explains why a group of organisms constitute a species should also be fundamental to the explanation of biological diversity.

The intellectual underpinnings of our definition can be traced to Goldschmidt (1940) and many others since. The recent contributions to evolutionary theory of Brooks and Wiley (1988), Wicken (1987), Colier (1988), and the Artificial Life Community (see, for instance, Langdon 1989, Kauffman 1992), treat some of the questions raised here.

Our variant of genomic definition is intentionally cast in the most general and informatic terms for the purpose of providing a definition able:

- i.) To encompass sexual and asexual species,
- ii.) To encompass both cellular organisms and viruses,

iii.) To encompass all nucleic acid-based self-replicating organisms, including protobionts based on RNA (RNA world), and

iv.) To provide a rational link between biology and computer science (artificial life community, computational physics).

Probably any of these goals taken separately is worthy of a separate theoretical pursuit. It is interesting that a single species concept, if correct, might encompass all four goals simultaneously.

Any new species definition is useful to the extent that it broadens the area of discussion of this open scientific problem. We believe the majority of other definitions to be correct as well within their specified areas of validity. Due to the lack of complete overlap, a comprehensive comparison of these with the genomic definition is not attempted. The proposed definition reflects the basic assumptions of molecular genetics.

DEFINITION AND ASSOCIATED PRINCIPLES

Our proposal for the definition of a biological species is as follows:

A species comprises a group of actual and potential biological organisms built according to a unique genome program that is recorded, and at least in part expressed, in the structures of their genomic nucleic acid molecule(s), having intragroup sequence differences which can be fully interconverted in the process of organismal reproduction.

The major premises on which the definition is founded are:

1. The unity of the living world is based on the tree of reproduction connecting all organisms. Starting with the first organism, this genealogy consists of reproduction cycles part of which is an imperfect replication of nucleic acids, i.e., an organism is any minimal physical system that belongs to the genealogy whose reproduction is based on nucleic acid replication.

2. The universal pattern of evolution is the multigenerational existence of groups of phenotypically similar organisms, i.e., species (Fig 1), in which each group has a unique founder group or, more rarely, a unique combination of founder groups (hybridization, symbiosis). This premise implies the impossibility of reversion; the newly formed group cannot produce the members of the founder group as progeny, nor members of any other unrelated group, i.e., there is a discontinuity, an unbridgeable gap, between organisms of different species, despite a common ancestor. It is not required that founder species be extinguished after giving birth to a new species. Also, this premise allows for a limited possibility that organisms of a given founder group or a specific combination of founder groups can start a given new group in several independent events during evolution.
3. Every individual organism belongs to one and only one species (Kitcher 1987), as

the first member of the group, as any of the other members, or as a barren mutant (Fig 2).

4. Identical sequences of entire genomes in two organisms can not code different instructions within the natural phylogeny based on nucleic acids. However, various genomic sequences can code identical instructions.
5. Starting from the first organism(s) onward in phylogeny, there is a stage in the reproduction process at which parts of the phenotype(s) [i.e., expression of genome program(s)] of the parent(s) are used as initial conditions for the triggering of life cycles of the first-generation offspring. Only some of the genomes of the offspring are compatible with "maternal" initial conditions, i.e., organisms encoded by these genomes can reach reproductive maturity and leave their offspring in a given environment. In some first-generation offspring, in much rarer instances, a genome acquires

differences and produces changed initial conditions for the second-generation offspring organism, which will not be capable of supporting its genome through reproduction if it were a revertant back to the genome(s) of its grandparent(s). In even rarer cases, a filial genome is not compatible with its own initial conditions, i.e., offspring could not have a genome identical to their parent(s).

6. Simultaneously living organisms of one species can exist physically isolated from each other, and some generations can be represented by a single living organism (Fig 2).

The basic assertion of the proposed definition is that the genome is the organizational level which accounts for the species characteristics given as premises by having a species-specific program. Any phenotype as derived from a genome has a lesser chance to give a correct picture of the species phenomena than the genome itself. In many cases, organisms of a single or closely related species are more different than organisms of evolutionarily more distant species when only separate characteristics are compared. Because genomes require

wider molecular systems of which they are part, one might claim that organisms themselves are the only level on which species can be described. However, if one holds that a component of an organism called genome determines species characteristics, this does not mean the other components of the system are not necessary, that the wholeness of organism is denied, or that interactions of organisms do not play a significant role. It only means that the genome is the most proximal cause for such behavior of the system.

In our definition, the phrase "genomic nucleic acid molecule(s), having intragroup sequence differences which can be fully interconverted" is the basis for recognizing one species from another. It follows from premises 4 and 5. Reproductive isolation is a specific phenotypical consequence of this principle and is valid for most cases concerning sexual organisms. The organisms possessing genome sequences that are potentially interconvertible (even if reproductively isolated) or that are nonconvertible with genome sequences of organisms of other species (even when the later are phenotypically very similar, and/or reproductively connected to the former) would constitute a species (see example in Fig 2). Thus, a reversible mutation responsible for reproductive isolation does not constitute the splitting of one species into two. However, this event can be the beginning of an accumulation of changes

that crosses the threshold of inconvertibility in subsequent generations.

Any vertical (mutation, deletion, duplication, translocation, transposition) or horizontal (sexual recombination, transfection, symbiotic merge) changes of genomic sequences in reproduction contribute to the generation of sequence differences. Some of these will lead to phenotypic organismal diversity within species or, very rarely, to the creation of a new species. However, the individual characteristics of genomic DNA within species can be repetitively produced starting from different parents, from clones in asexual species, etc.

The assumption that second-generation offspring can be revertants to the grandparental genomic sequence means that any organism within a species has a capacity to start a line, eventually giving an organism possessing a genomic sequence identical to that of any other member of the species. At least one pathway of interconversion is equivalent to reversion to the genome of a common ancestor and then proceeding with changes of this genome equivalent to those in the line leading from the ancestor to the reference organism (Fig 3). One must keep in mind that a potential for reversion to the grandparental genome is realized very rarely (especially in the case of complex genomes).

For the simplest organisms, protobionts and viruses, the totality of possible genomic sequences which are interconvertible could be defined already and thus the consequences of the genomic species definition shown in practice. The potential convertibility of total genomic sequences, which is very rarely realized in species with longer genomes, can be of a high practical value if correctly estimated from the limited knowledge of genomic sequences and their meanings (instructions). The rates of actual convertibility of defined parts of the genomic sequence, and the extent of allowable differences and the probability of their conversion could be measured and estimated, respectively. The elaborated principle of genome sequence convertibility outlined above can serve as a nearly universal basis for species recognition (Drmanac and Crkvenjakov, unpublished information). It can be a useful parameter for systematics as the natural and general level for identification and definition of species.

Which sequence changes could not be converted, thus representing a species split? Which instructions are encoded by these sequences? One possibility is a set of interdependent instructions whose probability of simultaneous change is close to zero, and partial changes lead to nonviable or very unadapted organisms. This explanation assumes that speciation

is a result of accumulation of changes which at one moment become practically irreversible or an occurrence of a very improbable complex change which, when it happens, has an infinitely small chance to be reversed (for example, if the chances of forward and back mutations are 10^{-6} , the probability of reversion is 10^{-12}). The probabilistic explanation requires an arbitrary decision (but based on existing knowledge), which probability is sufficiently small. More interesting types of irreversible changes are those which do not allow survival due to incompatibility in generational links. The evolution of a genome is inseparable from the evolution of a wider organized molecular system (cell, organism) of which it is a component. This fact prohibits the intentional conversion of genomic sequence only, or the transplantation of a genome in a cell of another species, causing in both cases nonviability of the organisms thus obtained.

One-locus and two-locus models of inconvertible change can be defined. For a one-locus model, let IC_x be a phenotype expressed as a part of initial conditions which is both encoded by locus L_x and activates L_x ; let $On(IC_x - L_x G_m)$ be successive organisms in the genealogy whose genomes are denoted by G_m , and whose IC are defined by parent genomes. Further L_1 encodes IC_1 , which can activate L_1 and L_2 ; L_2 encodes IC_2 , which can

activate L2 and L3 but not L1; IC3 can activate L2 and L3, but not L1. Then the simplified genealogy (one parent, one offspring) can be defined:

SPECIES 1|.....SPECIES 2.....
 $O_1(IC1-L1G1) \rightarrow O_2(IC1-L1G1) \rightarrow O_3(IC1-L2G2) \rightarrow O_4(IC2-L2G2) \rightarrow O_5(IC2-L3G3) \rightarrow O_6(IC3-L2G2) \rightarrow O_7(IC2-L2G2) \rightarrow$

Once created L2, (organism O_3) can not revert on L1, but L3 in organism O_5 can revert on L2 to give O_6 and O_7 , having genomes identical to that in O_4 .

Two-loci models requiring one change in each of two successive generations are also possible. In addition to the previous definitions, let Ty be target locus which is activated by initial condition factors encoded by Lx. Further, IC1 is compatible with T1 and IC2 with T2, but there is no cross compatibility; T1 activates expression of L1 and L2; T2 can activate expression of L2 but not L1. Then the genealogy can be defined:

spec1.....|.transient spec2|....spec3.....
 $O_1(IC1-L1T1G1) \rightarrow O_2(IC1-L2T1G2) \rightarrow O_3(IC2-L2T2G3) \rightarrow O_4(IC2-L2T2G3) \rightarrow$

Organisms IC2-L2T1G2 can not survive without changing T1 into T2. Neither can they revert back in G1 type. So it is important to note that both mutations can not occur in a single

generation. Note also that Tx can activate Lx indirectly through a regulatory chain, and that, like in the one-locus model, either Tx or Lx has to change two properties by acquiring one or more changes. Organisms having genomes of the G2 type represent a transient species.

Considering the inevitability of individual differences, what is the theoretical framework for discreteness and uniqueness of species? This framework is provided by the genome program as an integral set of instructions encoded in the sequence. Premise 4, that similar but nonidentical sequences can encode the same instructions, allows the possibility that a group of organisms can share a unique set of instructions. The program is understood as a specific pathway of self-reproduction of a particular material system, i.e., a specimen biological organism of a species. A species genome program is the product of ancestral programs and environmental constraints. It is a framework limiting the range of individual differences, adaptational possibilities, and capacities for starting a new species. The program is an entity which includes constraints which limit changes of both the record and the whole system and consist at least of the developmental subprogram (ontogenic program), reproduction subprogram, and the inscription rule, which is a specific way

of making the program record for development and reproduction. By the inscription rule we mean the class of phenomena of which an example is that one protein with the same time-space expression pattern can be coded in two species by genes with different numbers of introns and exons or even by genes without introns.

In our definition, species are considered as inevitable products of the process of self-reproduction of biological organisms. The formation of these discrete and enduring entities by self-reproducing systems in general and biological organisms in particular is the consequence of several irreducible properties of these systems. These properties are realized in the first self-reproducing system(s) (biological organism[s]) which was formed from nonreproducing material systems. Other, derived properties can be added on in organisms formed in an evolutionary chain, but the basic ones cannot be lost. The basic properties of self-reproducing systems are as follows:

- i) The wholeness of the organism, which makes it distinguishable from the environment and cannot be reduced to any one part of the system; and
- ii) The existence of a system part that has both the property of a structural element and the property of material record of a formal entity: a program for

self-reproduction. The system retains the capacity for self-reproduction within the range of changes of this part, and copying of this part in reproduction is generally imperfect.

- iii) The group of genealogically connected systems/organisms forms a species on the basis of the survival of a unique program encoded in their respective records which are imperfect but functional copies of the record of the first organism in that species.
- iv) The possibility exists of a birth of organism(s) of a new species (which may have similar, lower, or higher organismal complexity) from the member(s) of the old one(s) through consecutive changes of both the record and system in one or more reproductive cycles.
- v) The general tendency of new species is towards higher organismal complexity.

The wholeness of the system had to emerge once at the very beginning of the process, since its basic properties cannot be attributed to the system's components *per se* (i.e., when they are outside the system). The first system's component, which can not be regenerated outside the system after emergence (only within it with the help of other components), is the material

record of program. The system had to contain various accessory components in addition to the record component. These accessories are helpers which need not necessarily have been formed within a system (but could have been) and which are internalized within a system in the course of evolution.

The wholeness of the first biological system can be described in light of the dual properties of the nucleic acid molecule. This molecule provides the basis for replicable program inscription through chemical complementarity. Simultaneously, as a material component, it constitutes a part of the system and has a structural role, which in interaction with internalized components from the environment (accessory components) forms the whole of the reproducing organism. The structural role of nucleic acids has been retained in present-day organisms, as has the duality of structure and inscription.

Nucleic acids also satisfy the necessary property of self-reproducing systems of generally not requiring perfect copying of its program record. Replication based on complementary pairing is not entirely sequence context-specific in the sense that the fidelity of any base pair replication is mostly independent of neighboring sequences, but is dependent on some structure of either the replicating enzyme or the nucleic acid itself. Not only the sequence-derived structure can alter fidelity of replication and even prevent it, but also several

different sequences can give the same derived structural element, as evident in modern organisms. Complementary pairing opens the additional possibility of large and immediate sequence changes via recombination.

Generally, imperfect reproduction leads to organisms that differ from each other, but their genomic inscriptions are interconvertible, i.e., can be repeatedly generated from previous organisms. One can ask what is required, after sufficient time has elapsed, for the inevitable occurrence of organisms so changed that they, although being able to reproduce, are unable to revert to the range of parental types no matter how much time is allowed for it. Or, more directly put, what is the cause of existence of the species category. There are two aspects of this question. One is, why there is discreteness, i.e., gaps between species? The other is, why is the birth of an organism of a new species from an old species a rare and irreversible event? One can expect that certain levels of complexity of both system and environment are necessary to provide the sufficient conditions for organismal, and hence species, occurrences. This granted, the inevitability of species occurrences as entities separated by gaps is a consequence of the discrete nature of physical laws whose resultant derivative is a genome program. The gapped nature of our physical universe consisting of particles,

elements, stars, etc., makes this obvious once species are recognized as parts of its natural hierarchy.

The irreversibility and uniqueness are the result of a probability so low as to represent a virtual impossibility of reproduction in certain connecting stages through which systems would have to pass to both convert back and remain self-reproducing. An alternative concept sometimes mentioned is that speciation is a historical event and hence is non-repeatable. However, the spatio-temporal continuity is not a general property of species category, since exceptions exist.

DISCUSSION

The program is the primitive term (i.e., the meaning is assumed) of our definition, but we view it as a closed and finite information system or a heuristic algorithm. A system's program is a formal, logical entity and can be recorded by any material means to exist as a copy outside of the system simultaneously, before, or after it (Langdon 1989). Biological organisms are the only natural systems built according to a program. It is not known if other kinds of program-based systems exhibiting life-like properties are physically possible.

It is of interest to enlarge on the concept of the uniqueness of species. The concept emphasizes that a given species always arises from unique parent(s). Since the duration of parent species is fixed, the origin of new species must take place within this time span. Species are considered in literature either as spatio-temporal entities or as structural entities that can be repeated in various times. For ontological reasons, we prefer the structural view. Otherwise, one would have to deny the possibility of reappearance of extinct species (for example, the mammoth) or the formation of new ones through genetic engineering. Quite aside from the technical and ethical feasibility of such procedures, we are

concerned with the question of whether artificial organisms so formed would be a members of a given species if spatio-temporal continuity is required. Human agency notwithstanding, there is an almost perfect correspondence of natural species existence and their spatio-temporal continuity. Due to species complexity, it is very unlikely for them to recur. The need for structural definition of species is most readily seen at the point of emergence of phylogeny from genealogy. The fine topography of any speciation branching in phylogeny is currently impossible to observe or experiment upon. A requirement of spatio-temporal continuity is too restrictive for a general theory of speciation branching, since there is the possibility of merging of the populations of the same new species that originated from the same parent species at different times. The new species is structurally but not historically unique (an example is the lizard species that arises by hybridization of two parent lizard species and have done so apparently several times within the recent past).

Our basic assumption is that all organisms must belong to a species. Some variants of phylogenetic and ecological species definitions share this premise with the genomic definition. The biological species concept (Mayr 1942) introduced the alternative assumption that only sexual organisms must belong to species. Both assumptions are present

in the literature; the former as a basic unvoiced tenant in molecular genetics, the latter as a majority, but not unopposed, view in evolutionary biology. The controversial nature of the claim advanced here of species being as basic to biology as is the organism, points to the unfinished business of biology compared with other natural sciences, which (as presumably simpler) enjoy a fundamental conceptual unity. The fact that the program has evolved in some complex species to reside in more than one physically separable material system, for example, a breeding pair in sexual species, is not actually a major divide in evolution separating apparent and real species. As evident in many independent evolutionary lines, asexual species coexist with the sexual ones. This indicates that, at least in some cases, a relatively less profound genomic change than implied by the divide hypothesis is required to transform one into another.

Either of the two assumptions allows one to focus on the given pattern of grouping of organisms as true and in second step describe the process which is fundamental for generation of the perceived pattern. The phylogenetic definition of Nixon and Wheeler (1990) of species as "the smallest aggregation of populations (sexual) or lineages (asexual) diagnosable by unique combination of character states in comparable individuals (semaphoronts)" is interesting in this regard. It

shares with the genomic definition the first two of the outlined logical steps: the universality of species assumption and the perception of the pattern that for each real grouping of organisms must exist a set of shared character-states unique to the group. It does not identify the causative process responsible for the pattern. This is, of course, why this position is criticized from those within cladism who appear to accept the non-universality assumption and interbreeding as the basic process responsible for species generation (de Queiroz and Donoghue 1990). Our proposal of the inconvertibility of genome programs as a universal process, in addition, calls for a hierarchy of character states in deducing the process from the pattern. The sequence interconvertibility is a measure of the compatibility of sequence as a structure and a given genome program as a system of structure-function relationships. The same logical operations of analysis of congruence of structures and structure-function relations are used in the practice of systematics regardless of the actual definition espoused. While cladists might give equal weights to behavioral, morphological, and sequence character-states in case of their conflict, the genomic concept of species focuses on the sequence level as the one on which the enumeration of species-specific character-states will be least problematic, due to the causal proximity of the sequence to the genome program.

The examples of lambdoid phages and *Escherichia* and *Shigella* spp. illustrate the point. The shortest description of lambdoid phages is that they share a common genome program but not most of the sequence; their DNA can recombine in restricted areas, ensuring exchange of entire genes or gene groups. These modules have analogous functions, and various combinations with the correct number and order of modules give viable phages. P22 phage infects only *Salmonella* while λ grows on *Escherichia*. The two differ in having different head and tail proteins, i.e., different operon modules of similar size in their genomes for injecting phage DNA into cytoplasm. The horizontal gene transfer mechanisms (conjugation plasmids, etc.) ensure that either phage DNA can enter the cross-specific cytoplasm and there recombine with the host-specific phage. Presumably, P22 phage can revert into λ phage through natural process involving a number of generations with the appropriate recombination events with other lambdoid phages (Suskind and Botstein 1978). This property places all of them into a single species by our definition. It is highly improbable for the reversion to occur through point mutations or any other randomized genomic sequence change. This example illustrates that sequence change improbability does not mean systemic impossibility. *Shigella* species differ from other enteric bacteria by pathogenicity resulting from the ability to live

inside epithelial cells. The standard classification based on this phenotype places them in separate genus from species of the genus *Escherichia*. On the chromosome sequence level, some strains and species of *Escherichia* are more diverged among themselves than any of them are from *Shigella*. Nearly all phenotypic differences can be traced to a nonessential 250-kb *Shigella* plasmid (Sansonetti et al. 1982). *E. coli* transformed with this plasmid becomes *Shigella* in pathogenicity. Accordingly, all *Escherichia* and *Shigella* species and strains should be considered one species by the criterion of sequence interconvertibility, which implies a shared genome program. It is difficult to see how a similar analysis could be made on the basis of alternative species definitions.

The viewing of species by the concept of genome programs is broader than the conventional one, since it includes transient species in lines, if they existed, connecting a long-living ancestor/descendant pairs of species. The transient species can have a small number of generations. For any given pair of closely related but morphologically well-separated species, the nature of genomic changes required to transform one into another determines if one or more transient species are required or not. Our definition is neutral to this central question of speciation mechanisms except to predict that the question is not resolvable on the phenotypic level alone.

Two different inscriptions can generally represent either a single or two different genome programs. The inscriptions representing one program (i.e., within species) are more similar among themselves than they are to any given inscription of another program. Due to the inseparability of the program and its inscription, there must be a rule that transforms one into the other. The study of these rules valid for smaller or larger groups of organisms should be the subject of systematics. The example of viruses already illustrates the natural displacement of a systematics based on more derived phenotypic levels by one based on sequences. Furthermore, the potential of sequence-based systematics to achieve a higher degree of operability with the availability of the genomic sequence data is apparent as well to practitioners at the cutting edge of systematics of more complex organisms.

The turning of molecular biology towards the genome can be seen from the interest in various genome projects. The empirical fact that RNA molecules can catalyze RNA phosphodiester bond formation led to the proposal of a protobiotic "RNA world" (Sharp 1985, Gilbert 1986), now widely accepted among molecular biologists. Thus there is a reason now to consider only the option of RNA-based life, the first living system, which had a genome consisting of RNA and had no energy-transforming machinery. Curiously, the simultaneous

theoretical treatment of genome and the emergence of first organisms argued for here seem not to have been attempted previously.

Regarding our four goals, we have provided criteria of convertibility of genomic inscriptions that bond together the clones of asexual species as firmly as mating bonds the organisms of sexual species. By uniting the acellular and cellular species by the requirement that they both reproduce on the basis of a nucleic-acid-inscribed genome, the viruses, despite their lack of metabolic machinery, are shown to be a part of the living world. There is no real difference in the reproduction of a protobiotic RNA organism in primordial soup and of a phage infecting bacteria. Both genomes utilize the material and energy resources available in their respective environments. If one was alive by virtue of being ancestral to cellular life, the other must be too. The placement of a genome program in the center of biological species phenomena provides a theoretical basis for the already perceived need to expand the interactions between biology and computer science. The genome program as an algorithm lies in the realm of information sciences, and, by extension, in part in the realms of theoretical physics and mathematics.

The genomic species definition satisfies the requirements of philosophers by conforming to the three principles of

Kitcher (1987), as follows. Each organism belongs to exactly one species, since it is a system with one and only one genome program. All species taxa belong to the species category, since each has a unique genome program. Finally, the ultimacy principle is satisfied by the assertion that the genome level is the fundamental one for the explanation of biological diversity. This is based on the analysis of the case of the first organism/species proposed to have consisted of a replicating RNA molecule and accessory components. The laws valid for the first organism/species are the only laws common to all living beings descended from it.

ACKNOWLEDGMENTS

We thank Ivan Labat for helpful comments on the manuscript, David E. Nadziejka for editorial assistance, and Nikola Tucic for use of his library on evolutionary biology. An earlier version of this manuscript was completed at the Institute of Molecular Genetics and Genetical Engineering, Belgrade, Yugoslavia. The work was supported in part by the Science Fund of Serbia, Yugoslavia; U.S. Department of Energy grant DE-FGO2-88ER60699; U.S.-Yugoslav Joint Board grant JF-820; and by the U.S. Department of Energy, Office of Health and Environmental Research, under contract No. W-31-109-Eng-38.

REFERENCES

- Alberch, P., 1980, Ontogenesis and Morphological Diversification, *Am. Zool.* 20: 653-667
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D., 1989, *Molecular Biology of the Cell*, 2nd ed., Garland, New York
- Bodnar, J.W., Jones, G.S., and Elis, C.H., Jr., 1989, The Domain Model for Eukaryotic DNA Organization: A Molecular Basis for Constraints on Development and Evolution, *J. Theor. Biol.*, 137, 281-320
- Brooks, D.R., and Wiley, E.O., 1988, *Evolution as Entropy: Toward the Unified Theory of Biology*, 2nd ed., Univ. of Chicago Press, Chicago
- Colier, J., 1988, The Dynamics of Biological Order, in *Entropy, Information and Evolution: New Perspective on Physical and Biological Evolution*, B.H. Weber, D.J. Depew, and J.D. Smith, eds., MIT Press, Cambridge, pp 227-242
- Depew, D.J., and Weber, B.H., 1988, Consequences of Nonequilibrium Thermodynamics for the Darwinian Tradition, in *Entropy, Information and Evolution*, Weber, B.H., Depew, D.J., and Smith, J.D., eds. MIT Press, Cambridge, Mass. pp 317-354

- de Queiroz, K. and Donoghue, M.J., 1990, Phylogenetic Systematics and Species Revisited, *Cladistics* 6: 83-90
- Ehrlich, P., and Raven, P., 1969, Differentiation of Populations, *Science*, 165: 1228-1232
- Gilbert, W., 1986, The RNA World, *Nature*, 319:618
- Goldschmidt, R., 1940, *The Material Basis of Evolution*, Yale University Press, New Haven, Conn.
- Kauffman, S.A., 1969, Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets, *J. Theoret. Biol.*, 22: 437-467
- Kauffman, S.A., 1992, *Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford.
- Kitcher, P., 1987, Ghostly Whispers: Mayr, Ghiselin and the "Philosophers" on the Ontological Status of Species, *Biology & Philosophy* 2: 184-191
- Langdon, C.A., 1989, Artificial Life, in *Artificial Life*, Langdon, C. A., ed. Addison-Wesley, Reading, Mass., pp. 1-48
- Mayr, E., 1942, *Systematics and the Origin of Species*, Columbia University Press, New York.
- Mayr, E., 1970, *Populations, Species and Evolution*, Harvard University Press, Cambridge, Mass.
- Murtha, M.T., Leckman, J.F., and Ruddle, F.H., 1991, Detection of Homeobox genes in Development and Evolution, *Proc. Natl. Acad. Sci. USA*, 88: 11711-10715

- Nixon, K.C., and Wheeler, Q.D., 1990, An Amplification of the Phylogenetic Species Concept, *Cladistics*, 6: 211-223
- Ptashne, M., 1986, *A Genetic Switch*, Blackwell Scientific Publications, Palo Alto, Calif.
- Sansonetti, P.J., Kopecko, D.J., and Formal, R., 1982, Involvement of Plasmid in the Invasive Ability of *Shigella flexneri*, *Infec. Immun.*, 35: 852-860
- Sharp, P., 1985, On the Origin of RNA Splicing and Introns, *Cell* 42: 397-400
- Singer, M., and Berg, P., 1991, *Genes & Genomes*, University Science Books, Mill Valley, Calif.
- Suskind, M. M. and Botstein, D., 1978, Molecular Genetics of Bacteriophage P22, *Microbiol. Rev.*, 42: 385-413
- Tempelton, A., 1989, The Meaning of Species and Speciation: A Genetic Perspective, in *Speciation and its Consequences*, Otte and Endler, eds., Sinauer Associates, Sunderland, Mass., pp 3-27
- Waddington, C.H., 1957, *The Strategy of Genes*, Allen and Unwin Ltd., London
- Wesson, R., 1991, *Beyond Natural Selection*, MIT Press, Cambridge, Mass.
- Wicken, J.S., 1987, *Evolution, Thermodynamics, and Information: Extending the Darwinian Program*, Oxford University Press, Oxford

Wiley, E., 1981, *Phylogenetics: The Theory and Practice of
Phylogenetic Systematics*, Wiley & Sons, New York

FIGURE LEGENDS

Fig 1. The irreversibility of speciation in time. Each time horizon contains a certain number of phylogenetically related species (each represented by a different symbol). The irreversible founder/descendent relation is indicated by the same founder(s) in occurrence(s) in which a species appears as a progeny. Also present are different time-cohorts of the same species (identical symbols in two adjacent time-horizons) shown as either giving rise to different new species, or to the same one. Dashed arrows and symbols denote relations excluded by the irreversibility premise (see text). Some species (Δ , $\square$) are eliminated in the course of evolution.

Fig 2. Single species membership of organisms and origin of new species. Organisms belong either to species A or a new species B. In this example, the progeny of members of sexually compatible species A and B is always B, as required by the irreversibility premise. Any two A organisms are capable of giving A progeny. Only a limited set of A $\times$ A pairs gives rise to B

progeny. $B \times B$ crosses give B only, as well as $A \times B$ crosses when viable.

Fig 3. Genome interconversions within species phylogeny. Member organisms (O) have different subscript numbers to denote their nonidentity as physical objects. However, their genomes (G) can be identical in sequence to ancestral genomes. O_1 , O_6 and O_{13} contain G_1 , O_2 , O_9 and O_{10} contain G_2 , O_4 and O_{14} contain G_4 , and O_{11} and O_{15} contain G_8 , respectively.

Fig 2

Fig. 2

Fig 3

DATE

FILMED

7/13/94

END