Object Spaces: An Organizing Strategy for Biological Theorizing

Beckett Sterner
Conceptual and Historical Studies of Science
University of Chicago
Chicago, IL, USA
bsterner@uchicago.edu

Abstract

A classic analytic approach to biological phenomena seeks to refine definitions until classes are sufficiently homogenous to support prediction and explanation, but this approach founders on cases where a single process produces objects with similar forms but heterogeneous behaviors. I introduce *object spaces* as a tool to tackle this challenging diversity of biological objects in terms of causal processes with well-defined formal properties. Object spaces have three primary components: (1) a combinatorial biological process such as protein synthesis that generates objects with parts that are modular, independent, and organized according to an invariant syntax; (2) a notion of "distance" that relates the objects according to rules of change over time as found in nature or useful for algorithms; (3) mapping functions defined on the space that map its objects to other spaces or apply an evaluative criterion to measure an important quality, such as parsimony or biochemical function. Once defined, an object space can be used to represent and simulate the dynamics of phenomena on multiple scales; it can also be used as a tool for predicting higher-order properties of the objects, including stitching together series of causal processes. Object spaces are the basis for a strategy of theorizing, discovery, and analysis in biology: as heuristic idealizations of biology, they help us transform inchoate, intractable problems into articulated, well-structured ones. Developing an object space is a research strategy with a long, successful history under many other names, and it offers a unifying but not overreaching approach to biological theory.

Keywords

causal process, combinatorics, fitness landscape, heuristics, morphospace, sequence space

October 31, 2008; revised and accepted October 31, 2009
Biological Theory 4(3) 2009, 280–286.

Not all causes are equal. A primary challenge for biology is to identify the causes of important phenomena, but studying the causes separately often produces disconnected models of varied form, scope, and utility. One might wonder, then, whether a more cohesive and general theory is possible for biology? No grand unified theory or quantum Standard Model seems appropriate to evolution's complexity and exceptionalism. A different approach, however, proceeds under the perspective of understanding the simultaneous diversity and common origin of life in terms of a number of causal processes that combinatorially assemble a small alphabet of parts into a large set of products. These combinatorial processes thread together the otherwise baffling variety of biological objects, such as proteins, species, and organs, because all processes share a common formal structure. I introduce the term *object space* to refer to a combinatorial process, the set of possible objects it can produce, and the analytical tools we have for studying the objects' behavior. Developing an object space is a research strategy with a long, successful history under many other names, and it offers a unifying but not overreaching approach to biological theory.

A common alternative theoretical approach hypothesizes that biology is structured in hierarchical levels: lowest are molecules, then moving upward are organelles, cells, organs, organisms, ecosystems, etc. A general theory for a given level would model the logic of that level's constituents' interactions and their causal connections with other levels. Even within the level's hypothesis, however, there are substantial problems that arise from the heterogeneity of supposedly similar objects. Accommodating various objects that are all on the same level in one theory can prove intractable. For instance, there may be little theoretical structure shared between models of desert, rainforest, and tidal pool ecosystems. Alternatively, a single model for all cells may founder on how some cells cannot live on their own, others move in and out of communities, and still others always stay independent.

Such heterogeneity often provokes the strategy of reexamining a confounding object type for possible reclassification into several new, more homogenous categories. This analytic strategy is classic and proven, but as with all strategies, it succeeds only under particular conditions. One problematic situation occurs when the objects have little in common once they exist but come about through a shared mode of generation. In this case, a single class is still justified, but models of behavior alone will produce isolated, even contradictory, results. This case therefore deserves a complementary strategy, which also has an illustrious history but has not yet been fully articulated.

This alternative strategy seeks to discover how the objects are assembled modularly through a common combinatorial process. When appropriate, this strategy uncovers the organization of the assembly process and thereby provides a recipe for generating theoretical knowledge of all of the process' possible products and their properties. Considered all together, the results are what we can call an *object space*.

Common types of object spaces in biology include sequence space for DNA, RNA, and proteins; the space of possible phylogenetic trees; morphospace; the conformation space of protein bond angles; and the interaction network space for enzymes or genes. As is clear, these examples have provided important tools for analysis and prediction, but to date their shared conceptual underpinnings have received relatively little attention.

Three main properties define object spaces. First, the common combinatorial process generates objects whose parts are modular, independent, and organized by an invariant syntax. Second, the space defined by all possible products of the process and its syntax often also carries rules for dynamic change over time as found in nature or used by algorithms. Third, functions defined on the space map it to other spaces or apply evaluative criteria to measure a desired quality, such as parsimony or biochemical function. In other words, the combinatorial process generates a set of objects with ordered parts, the notion of change implies a sense of "distance," and mapping functions establish a relation between different kinds of objects. Stadler et al. (2001) give an excellent exposition of mathematical forms of distance appropriate to mappings between genotype and phenotype space for RNA molecules, and Mitteroecker and Hutteger (2009) discuss the geometries and limitations of Euclidean distance intuitions for morphological spaces.

A good summary of object spaces' usefulness is that they help us to articulate well-structured problems. Simon (1973) listed several requisites to make a problem straightforward, including a test for proposed solutions, a representation for the initial, final, and intermediate system states, and rules for how to change between states. Any object space provides a powerful range of well-structured problems; for example, protein sequence alignments require that sequence comparison be a well-structured problem. Simon goes further to argue that even ill-structured problems, such as designing a new house, can be tackled by decomposing the problem into many well-structured problems and combining the results using heuristic methods. The problem-solving model he describes suggests that developing our knowledge of object spaces can provide a cohesive and general strategy for studying the complexity of biological phenomena.

Two Examples: Sequences and Phylogenetic Trees

Before we address the three qualities of object spaces, two historical examples merit further detail. The history of their introduction and justification can illuminate how to apply a similar strategy in the future. Also, the examples can serve as concrete examples of the three qualities to be discussed.

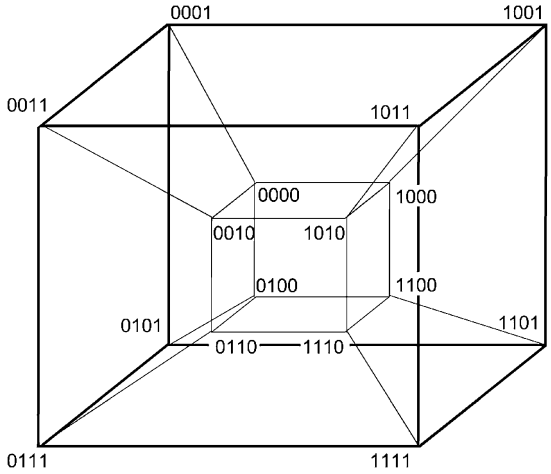

Figure 1.
A hypothetical, 4-dimensional sequence space where each position can be 0 or 1. The space is a hypercube with sequences on each corner and edges representing the action of flipping a position's value (Kauffman 1993).

It is impossible to do justice to their science and history in such a short space, so I will only sketch a couple of the most important aspects.

Both sequence and phylogenetic tree space share their origins in the discovery of DNA and Crick's (1958) paradigmatic exposition of the Central Dogma. The Dogma's main tenet—that information only flows outward from DNA to proteins and not in reverse—justified the expectation that knowing an organism's DNA should largely determine everything else about it. The breaking of the genetic code in the mid 1960s made clear how cells synthesize proteins from DNA: Every three DNA nucleotides translate into one amino acid, and the 64 sets of three nucleotides map redundantly to the 20 amino acid types. Two more discoveries—that ribosomes synthesize proteins linearly one amino acid at a time, and that many proteins fold on their own in solution—combined with the Central Dogma and the genetic code to justify efforts to predict the structure and function of proteins from their DNA sequence alone.

The coining of sequence space followed shortly (Maynard Smith 1970; Kauffman 1993). There are at least two definitions of sequence space, but the most common one is defined for DNA as follows: DNA molecules can be represented as a linear, ordered sequence of nucleotides A, C, G, and T. Each nucleotide position in the sequence corresponds to a dimension in the space, so a sequence that is N positions long implies an N-dimensional space. Since each dimension has four possible states, there are $4 \times N$ points total, so each point represents a sequence of N nucleotides. The high dimensionality of sequence spaces make visualizing them difficult—a typical feature of object spaces—but Figure 1 represents a space for a hypothetical sequence of length 4 with two states 0 and 1 for each dimension.

One major use for sequence spaces is to represent evolutionary change as DNA mutations move genes from one point to another. Zuckerkandl and Pauling (1965: 98) claimed a major advance for evolutionary theory when they proposed chemical paleogenetics as the comparison of DNA sequences that descended from a common ancestor. They argued that evolution could be represented as a phylogenetic tree whose nodes were DNA sequences, and that the new representation was superior because both causation and form coincided in DNA:

> In order to accept the special importance of the analysis of informational macromolecules, it is sufficient to subscribe to the following propositions: (a) The level of biological integration that contains the greatest concentration of "causal factors" will further our understanding of life more than any other. (b) A concentration of information is a concentration of "causal factors." (c) The largest concentration of information present in an organism, and perhaps also the largest amount of information, and the only organically transmissible information, is in its semantides [i.e., DNA, RNA, and proteins]. (Zuckerkandl and Pauling 1965: 98)

This argument of their seminal paper in phylogenetics and molecular biology highlights the connection we will see below between common combinatorial processes and the existence of a general representation for an object space.

Zuckerkandl and Pauling (1965) clearly advocated a genetic reductionist strategy, and it is important to pause and emphasize that my conception of object spaces requires no reductionist commitment. There are at least three senses in which object spaces are an open-minded strategy. (1) Developing an object space is a heuristic, in Wimsatt's (2007) sense of a reasonable plan without formal guarantees of success given the large uncertainties at work. (2) Object spaces depend on empirical evidence to characterize the combinatorial process, and not on *a priori* reasoning. (3) Object spaces match the flexibility of evolution's constant differentiation and radical change. Using object spaces will likely remain closer to the heuristic of designing a new house by decomposing it into solvable subproblems instead of applying a deductive framework of laws. Moreover, although many of the spaces listed as examples are molecular in focus, morphospace and phenograms in quantitative taxonomy indicate the broad, nondogmatic potential of object spaces.

Returning to phylogenetic tree space, Zuckerkandl and Pauling (1965) helped initiate decades of research aimed at inferring the history of evolution by sampling a number of DNA sequences and applying an evaluative criterion to decide which phylogenetic tree best explained the observed sequences. Phylogenetic trees are diagrams that represent the branching process of evolution: over time, the ancestral populations, represented by the higher nodes in the tree, branch downward into descendant sequences, ultimately forming the

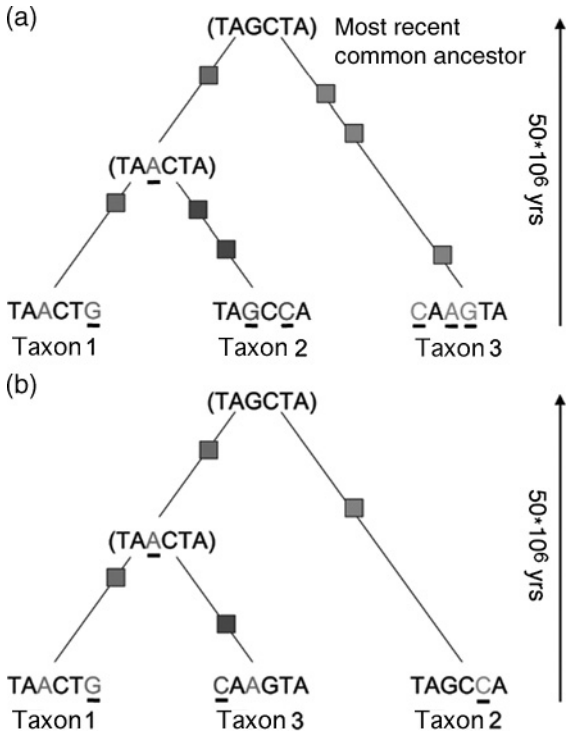

Figure 2.
Two phylogenetic trees of DNA sequences related by switching the positions of taxa 2 and 3. Shaded squares mark historically inferred mutations, and dark lines underneath nucleotides indicate which amino acids changed since the most recent common ancestor. Note that switching the taxa in tree A decreases the number of inferred mutations.

bottom-most nodes observable today. Phylogenetic trees vary in two main dimensions given a fixed number of sequences: where branching events occur in the tree, and where sequences are located on the nodes. There is no easy way to visualize all the trees at once in "space," so typically the possibilities are simply listed side by side (see Figure 2).

We should pause to note that although the focus of this discussion is the common structure of object spaces on a high level of abstraction, in practice scientists would relate to object spaces differently depending on how concretely they work with the represented objects. A scientist working to predict protein structures from sequences will work mainly with the models that link these two spaces, not the more abstract properties of spaces themselves, which remain in the background. A scientist studying a regulatory mechanism will rarely need even the great generality of sequence space because only a handful of mutations of that sequence are relevant. This article will be helpful to more concrete problems mainly when they can be solved by generalizing upward instead of gathering more detail.

The Combinatorial Process

A combinatorial process joins together a small number of kinds of modules in a particular order to generate a set of products many orders of magnitude larger in size. Examples of combinatorial processes can involve physically concrete mechanisms such as protein synthesis, or abstract mechanisms such as the branching process of speciation in evolutionary trees. In the former case, the process generates a physical object out of other physical objects; in the latter evolution, as we understand, it generates phylogenetic trees as historical facts out of branching and mutation events. The most important property of a combinatorial process is that its products can be represented as parts combined using a single syntax. The syntaxes of protein sequences and phylogenetic trees are straightforward: Amino acids form an ordered list without branches or loops, and trees are nested hierarchical pairs without cycles.

Some combinatorial processes are sufficiently vague as to impose no relative ordering on their components, resulting in a *degenerate* object space. Degeneracy is a term from physics (not a pejorative) meaning that no relative ordering differentiates the space's dimensions, which therefore can be interchanged without changing the objects' properties. Degeneracy occurs in two major cases: in the first, "realist" case the combinatorial process itself imposes no relative order, so that the objects' syntax is merely an unordered list of properties; in the second, "nominalist" case, we presume that an ordered, generative process exists but we describe it in terms not assumed to reflect the process's true structure. The first case is common for object spaces in physics, such as the state space of a gas where its pressure, temperature, and volume are invariant when the velocities or locations of any two molecules are exchanged.

The second case is common when we know more about the end states of a causal process than how it actually happens. David Raup's theoretical morphospace for invertebrate coiled shells (the kind you find on a beach) is a good biological example. Raup and others developed a mathematical formula with four parameters that could generate the range of possible shell geometries (Raup 1966). Defining the parameters as the axes of the morphospace, Raup examined which parts nature had left unoccupied and why. Raup's formulas model only the macroscopic aspects of shell growth, and the model's abstract parameters do not correspond directly to factors in the developmental process. The morphospace is degenerate because the presumed developmental process does not necessarily correspond to the parameters in the abstract growth "process" of the formula and therefore imposes no syntactical ordering on them. So this second form of degenerate object space lacks a realistic alphabet but still presumes existing characters that combine in a structured way to produce the observed morphology. Brakefield and Roskam's (2006) recent work on empirical morphospaces offers another example that is closely tied to investigating the underlying developmental processes.

A hallmark of *non*-degenerate object spaces is that their syntax renders the space impenetrable to classical laws and abstract models, and instead requires computational heuristics. The need for computational methods is linked to the difficulty of formulating universal biological laws, but some object spaces in physics and chemistry also show similar complexity. Thus, while we can study the generation of small- to medium-sized atoms from subatomic particles using differential equations, the space of all molecules produced by combining a few chemical reaction mechanisms exceeds our mathematical abilities. The problems are complex because the objects' higher-order properties depend sensitively on the states of its constituent parts on the local, intermediate, and global scales. Mutating a single amino acid may stop a protein from folding, while changing the protein's net charge can disrupt its function as a binding agent. Stadler et al. (2001) have developed a topological and statistical perspective that characterizes the intra-dependencies of RNA sequences using rigorous mathematical terminology.

Rules of Motion and Invariance

With the combinatorial process and the resulting syntax defined, there are two major classes of problems that object spaces help to tackle. The first concerns dynamics and algorithmic methods that depend on rules of motion in object space. The second, addressed in the next section, concerns prediction and analysis using functions defined on the space that map it to other spaces.

Sequence space is a good example for defining motion in object spaces and their theoretical interest for modeling evolutionary dynamics. A few general mechanisms cover most of the ways sequences change over time: one DNA nucleotide is replaced by a different one, a nucleotide is inserted or deleted, or whole groups may be changed by recombination or inversion. These mechanisms define rules for moving between sequences and imply a range of possible distance measures. One common measure is edit distance, which simply counts the number of moves needed to get from sequence A to sequence B. More sophisticated distance measures incorporate redundancies in the genetic code and empirical observations about varying amino acid replacement frequencies.

As an aside, the existence of multiple distance measures should indicate that many ways of measuring relative location could apply to a single combinatorial process. In fact, many spaces are defined by mathematically weaker notions of nearness or neighborhood, so even a traditional idea of Euclidean distance is not essential to a space. In general, the context of the biological situation will define the appropriate notion of distance, and this contextual aspect of object spaces is another reason biology does not lend itself to universal biological laws. In a sense, organisms use the spaces to represent their environment, and it would be backwards for us to use the spaces to represent all organisms.

Continuing on, if we assign a rate to motion in sequence space, we can then study the dynamics of populations of sequences—the molecular clock hypothesis states that nucleotide mutations occur at a constant rate. While originally the hypothesis claimed a common rate for all organisms, it is now typically limited to short periods of time with slow environmental change or gradual evolution. Nonetheless, the hypothesis is a basic tool for modeling genetic evolution.

Adding time to object spaces produces one of its most useful aspects: simulation. For sequence space, we can simulate branching events in a phylogenetic tree by populating the space with sequences that mutate, reproduce, and die. Alternatively, Raup and Gould (1974) used a simulation of randomly generated evolutionary trees to test if seemingly selected characters may be indistinguishable from noise.

Dynamic modeling in object spaces can also represent multiscale phenomena. In sequence space, we can simultaneously model events for individuals, genetic drift and selection in populations, and the generation of higher taxa as populations diverge. Although each of these phenomena has received attention and theorization individually, they can be studied jointly in sequence space. Researchers are currently developing another example of multiscale dynamics with gene interaction networks. As more expression data and genomes become available, we can model the fluctuating interactions of genes across whole cells and even mutualistic communities of bacteria (Stolyar et al. 2007). Object spaces therefore facilitate theorization across classic "levels" of phenomena.

Often, however, we may want to abstract away the details of nature, either to simulate a process faster or to understand which factors matter. For example, protein folding poses a frustrating challenge to biophysicists: In nature, proteins usually fold in microseconds, yet even simplified simulations take hours, days, or months depending on the problem. Instead of mimicking nature, biophysicists often rotate whole segments of the protein at once in order to better sample conformation space or sometimes replace each amino acid in the model with a large spherical particle for easier analysis.

Even these rules of motion are connected to natural processes, but for some purposes, we may simply want to best satisfy some criterion. Because the total space is usually too large to search exhaustively (the number of objects grows exponentially with each new part added), one must apply heuristics. Many heuristic methods work by guessing initially at a best object in the space and then iteratively improving this according to search rules. (Imagine guessing tree A in Figure 2 and then swapping taxa 2 and 3 to minimize the number of mutations.) Their search methods do not happen in nature; rather, the

artificial rules of motion only optimize the evaluative criterion efficiently.

Mapping Functions

The second major use for an object space relates it to other spaces or theoretical concepts via a mapping function (see Stadler et al. 2001 for the mapping of genotypic space to phenotypic space). Mapping one space to another is a way of joining together two combinatorial processes: mapping sequence space to protein conformation space, i.e., the protein-folding problem, combines protein synthesis and folding. In addition, simulating the dynamics of objects may depend on determining their behavior with a governing evaluative function. Evaluative functions express a criterion in the form of a model or formula that maps each position in the space to a number representing how well it satisfies the criterion. In the multiscale simulation of evolution mentioned above, the evaluative function defines the fitness of each object at that time, creating a landscape in the space within which the sequences move. In sum, mapping functions can transform one space into another or map a given space onto the real numbers for the purposes of simulation or evaluation.

Mapping functions complement the unifying capabilities for biological theory of multiscale dynamics. Multiscale dynamics represent multiple processes in one object space, and mapping functions connect object spaces to other spaces or theoretical concepts. For examples of mapping functions, one need only look to much of the current research in computational biology, molecular biology, and bioinformatics that focuses on how one kind of object becomes another: How does a protein sequence determine its structure? How does a structure determine its function? How do interacting structures determine biochemical networks and reaction rates?

A potential benefit of object spaces as a research strategy is an increasing ability to articulate formally the contribution of the environment to biological processes. Object spaces complement ongoing efforts at theorizing causal or informational environmental contributions as being on par with genes (Griffiths and Gray 1994; Oyama 2000). The possibility of combining multiple object spaces, whether consecutively by mapping, perspectivally via scaling, or in parallel using standalone models, holds the promise of uniting multiple phenomena with radically different representations under a single, integrated computational framework (Stein 2008).

Conclusion

As a research strategy, developing object spaces can be organized by their three main qualities. The first step is identifying a causal process that combines a relatively small number of components into a larger object of combinatorial complexity. The process should allow variation in the organization of parts in the object while obeying a fixed syntax. Next, try to find rules of motion or relations of distance that describe how the objects vary over time, e.g., mutating genes in sequence space. Finally, uncovering selective forces or further downstream causal processes allows one to define mapping functions on the space that either transform the objects into objects in another space (e.g., sequences into folded proteins) or evaluate them by some criterion (e.g., fitness or enzymatic rate for a given chemical mechanism). Once defined, an object space can be used to represent and simulate the dynamics of phenomena on multiple scales; it can also be used as a tool for predicting higher-order properties of the objects, including mapping the given space to one that follows it in a larger causal process. Research to uncover and develop object spaces in biology has proven very successful, and much current research fits well under that heading.

Object spaces are integral to the study of biological processes. They help transform what initially appears impossible into a tractable problem subject to well-known heuristics and methods. Simon's (1973) characterization of how an architect tackles designing a new house is appropriate to describe how biologists tackle explaining the complex traits exhibited by any organism: even if the problem is ill-structured and admits no direct solution, it can still be decomposed into well-structured subproblems whose answers are recombined to approximate the original whole. Object spaces offer a multitude of well-structured problem spaces for the big questions in biology, and while they will not render biology well structured overall, they deserve a central place in biological theory.

Acknowledgments

Many thanks to Bill Wimsatt, Matt Haber, Chris DiTeresi, Elihu Gerson, Julio Tuma, Erin Barringer, and the University of Chicago Philosophy of Biology and CHSS graduate workshops. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship.

References

Brakefield PM, Roskam JC (2006) Exploring evolutionary constraints is a task for an integrative biology. American Naturalist 168: S4–S13.

Crick FH (1958) On protein synthesis. Symposia of the Society for Experimental Biology 12: 138–163.

Griffiths PE, Gray RD (1994) Developmental systems and evolutionary explanation. Journal of Philosophy 91: 277–304.

Kauffman SA (1993) The Origins of Order: Self-Organization and Selection in Evolution. New York: Oxford University Press.

Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225: 563–564.

Mitteroecker P, Huttegger SM (2009) The concept of morphospaces in evolutionary and developmental biology: Mathematics and metaphors. Biological Theory 4: 54–67.

Oyama S (2000) The Ontogeny of Information: Developmental Systems and Evolution, 2nd ed. Durham, NC: Duke University Press.

Raup DM (1966) Geometric analysis of shell coiling: General problems. Journal of Paleontology 40: 1178–1190.

Raup DM, Gould SJ (1974) Stochastic simulation and evolution of morphology: Towards a nomothetic paleontology. Systematic Zoology 23: 305–322.

Simon HA (1973) The structure of ill-structured problems. Artificial Intelligence 4: 181–201.

Stadler BMR, Stadler PF, Wagner GP, Fontana W (2001) The topology of the possible: Formal spaces underlying patterns of evolutionary change. Journal of Theoretical Biology 213: 241–274.

Stein LD (2008) Towards a cyber infrastructure for the biological sciences: Progress, visions and challenges. Nature Reviews Genetics 9: 678–688.

Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA (2007) Metabolic modeling of a mutualistic microbial community. Molecular Systems Biology 3: 92.

Wimsatt WC (2007) Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality. Cambridge, MA: Harvard University Press.

Zuckerkandl E, Pauling L (1965) Divergence and convergence in proteins. In: Evolving Genes and Protein (Bryson V, Vogel HJ eds), 97–166. New York: Academic Press.