A Partial Solution to the C-Value Paradox

Jeffrey M. Marcus

Department of Biology, Western Kentucky University,
1906 College Heights Boulevard #11080, Bowling Green KY 42101-1080
jeffrey.marcus@wku.edu

Abstract. In the half-century since the C-value paradox (the apparent lack of correlation between organismal genome size and morphological complexity) was described, there have been no explicit statistical comparisons between measures of genome size and organism complexity. It is reported here that there are significant positive correlations between measures of genome size and complexity with measures of non-hierarchical morphological complexity in 139 prokaryotic and eukaryotic organisms with sequenced genomes. These correlations are robust to correction for phylogenetic history by independent contrasts, and are largely unaffected by the choice of data set for phylogenetic reconstruction. These results suggest that the C-value paradox may be more apparent than real, at least for organisms with relatively small genomes like those considered here. A complete resolution of the C-value paradox will require the consideration and inclusion of organisms with large genomes into analyses like those presented here.

1 Introduction

In the years following the discovery that DNA was the hereditary material [1], and even before the structure of DNA was fully understood [2], investigators measured the amount of haploid DNA (or C-value) in the cells of various organisms, hoping that this quantity might provide insights into the nature of genes [3]. They found no consistent relationship between the amount of DNA in the cells of an organism and the perceived complexity of that organism, and this lack of correspondence became known as the C-value paradox [4].

The C-value paradox has become one of the enduring mysteries of genetics, and generations of researchers have repeatedly referred to the lack of correspondence between genome size and organismal complexity [3, 5-9]. In spite of all of the attention devoted to the C-value paradox over more than five decades, there has yet to be an explicit statistical correlation analysis between measures of genome size and measures of organismal morphological complexity. Organismal complexity has been difficult to examine rigorously because of the inherent difficulties in measuring the complexity of organisms. Rather than trying to measure morphological complexity, most researchers studying the C-value paradox referred explicitly or implicitly to a complexity scale with bacteria at the bottom; protists, fungi, plants, and invertebrates in the middle; and vertebrates (particularly humans) at the top (Figure 1). This scale, called the Great Chain of Being, can be traced back to Aristotle and exerts a pervasive

A. McLysaght et al. (Eds.): RECOMB 2005 Ws on Comparative Genomics, LNBI 3678, pp 97-105, 2005

influence over popular and scientific perceptions of complexity [10]. This scale is not quantitative and incorporates untested assumptions about the relative complexity and monophyly of taxonomic groups.

Instead, researchers have focused on studying the statistically significant relationships between genome size and a variety of other quantitative traits such as cell volume, nuclear volume, length of cell cycle, development time, and ability to regenerate after injury [5, 11-15], which had long been included as part of the C-value paradox [3, 4], but which recently have collectively been redefined as a distinct phenomenon known as the C-value enigma [6]. There have been two general categories of hypotheses concerning the cause of the C-value enigma and of genome size variation [5, 6]. Explanations in the first category suggest that the bulk of the DNA has an adaptive significance independent of its protein-coding function. The amount of DNA may affect features such as nuclear size and structure or rates of cell division and development, suggesting that changes in genome size may be adaptive [6, 12, 13]. The second category of explanation suggests that the accumulation of DNA is largely nonadaptive, and instead represents the proliferation of autonomously replicating elements that continue to accumulate until the cost to the organism becomes significant [16, 17]**.** Gregory [6, 14, 15] has recently summarized the available data and argues that variation in genome size (and by implication variation in amounts of genomic heterochromatin) is predominantly due to direct selection on the amount of bulk DNA via its causal effects on cell volume and other cellular and organismal parameters.

The work described in this paper is explicitly not an examination of the C-value enigma, which is relatively well studied. This study attempts to address a very different question, the C-value paradox *sensu stricto* or the relationship between genome size and organismal morphological complexity, which is virtually unstudied. There are several developments that have increased the tractability of this type of investigation. One of the most important of these is the availability of many organisms with sequenced genomes, providing us with reliable estimates of both genome size and number of open reading frames (an estimate of gene number) [18]. A second advance has been the development of measures of non-hierarchical morphological complexity [19]. The number of cell types produced by an organism is among the most commonly used indices of non-hierarchical morphological complexity, and there are cell type counts available for a wide variety of organisms [20-24]. A final advance has been the development of comparative techniques such as phylogenetically independent contrast analysis [25-27]. Phylogenetically independent contrasts allows the study of correlations among traits between different species of organisms, even though the organisms vary in their degree of relatedness and are therefore not independently and identically distributed. It does this by using an explicit phylogeny to create a series of contrasts between pairs of sister taxa which, by definition, are the same age, so the time elapsed and the accumulated phylogenetic distance between the sister taxa is factored out of the analysis. The resulting contrasts are independently and identically distributed, and therefore suitable for correlation analysis. The novel approach presented here builds on these developments, using measures of genome size and complexity from sequenced genomes, numbers of cell types and numbers of subcellular parts as measures of morphological complexity, and phylogenetically independent contrast analysis to provide the first explicitly statistical analysis of the C-value paradox. The results of independent contrast analysis suggest that the C-value and measures of morphological complexity are significantly positively correlated.

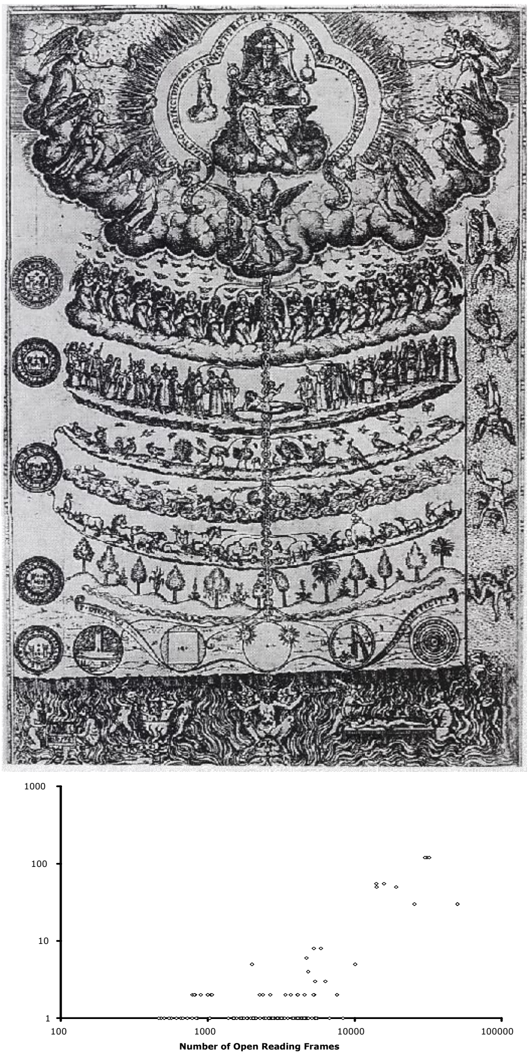

Fig. 1. Representations of the relative complexity of organisms. Above: Pictorial representation of the Great Chain of Being as depicted in Valades [28] (after Fletcher [29]). Below: the correspondence between number of open reading frames (an estimate of gene number) and the number of cell types produced for the first 139 organisms with sequenced genomes

2 Materials and Methods

While genome size has been the preferred metric for comparison with complexity, at least initially it was intended to be a proxy for gene number [3], which was difficult to estimate accurately until whole genome sequencing became possible. Genome size and number of open reading frames (an estimate of gene number) for the first 139 completely sequenced genomes (including 121 prokaryotes and 18 eukaryotes) were obtained from two genome databases [30, 31]. The number of cell types and for prokaryotes, the number of types of cell parts, produced by each organism was obtained from the literature (On-Line Supplementary Table 1). Cell types were considered distinct if intermediate morphologies were rare or absent. Counts of types of cell parts were determined from descriptions of prokaryote ultrastructure that were included in species descriptions. The number of cell types and the number of types of cell parts represent non-hierarchical indices of complexity—organisms with more cell types or cell parts are considered to be more complex than organisms with fewer cell types or cell parts [19]. To control for evolutionary relatedness that might confound correlations between these measures, I used independent contrast analysis with phylogenetic trees generated from the small subunit of ribosomal RNA [32], using sequences available for each taxon from NCBI [33]. Sequences were aligned in CLUSTALX [34] and phylogenetic trees were generated by Neighbor-Joining, Parsimony, and Maximum Likelihood methods as implemented in PAUP* [35] with Eukaryote 18s rRNA sequences used as the outgroup.

In addition, a data set for 45 taxa (4 eukaryotes and 41 prokaryotes), compiled and aligned by Brown et al. [36], and consisting of amino acid sequences for 23 conserved genes was kindly provided by James R. Brown. Open reading frame counts for two of the species included in Brown et al. [36], *Porphyromonas gingivialis* and *Actinobacillus actinomycetemosomonas* were not available when these analyses were conducted, so these species were excluded in my analysis (leaving 43 taxa: 4 eukaryotes and 39 prokaryotes). The Brown et al. (2001) data set was analyzed using Neighbor-Joining and Parsimony techniques. Jukes-Cantor branch lengths were applied to all trees, branch lengths of 0 were converted to 0.000001, and all branch lengths were transformed to the square root of the Jukes-Cantor distance to standardize them for analysis by contrasts [27].

3 Results

The evolutionary trees produced by the phylogenetic analyses are not shown because they largely replicate the results of Nelson et al. [37] and Brown et al. [36]. These analyses had to be repeated because independent contrast analysis requires that the species included in the phylogeny and the species included in the continuous character data sets must be completely congruent. Independent contrast analyses using trees produced by different tree-building algorithms from the same data set produced highly similar correlation coefficients, while analyses using trees derived from different data sets had larger differences in correlation coefficients. However, the significance of independent contrast correlations was generally robust to changes in tree topology, suggesting that phylogenetic uncertainty due to tree differences is unimportant in interpreting these correlations [38].

3.1 Small Subunit rRNA Phylogeny Independent Contrast Analysis

First, the small subunit rRNA data set will be considered. After independent contrast analysis, the correlation between number of cell types and genome size was significant or nearly significant depending on the tree used (Table 1), and the correlation between the number of cell types and the number of open reading frames was highly significant (Figure 1). As eukaryotes, with their larger genome size and greater number of cell types, might unduly influence these results, so the eukaryotes were pruned from the trees and the independent contrast analysis was repeated. A significant correlation was detected between the number of cell types and genome size, as was the number of cell types and number of open reading frames. To answer the concern that prokaryote cell diversity might be better expressed in terms of numbers of cell parts (organelle-like structures: prokaryotes by definition do not have true organelles), rather than numbers of cell types, the number of types of cell parts for each prokaryote was also collected. This yielded a significant correlation between the number of types of cell parts and genome size and between the number of types of cell parts and the number of open reading frames.

Table 1. Independent contrast analyses for the small subunit rRNA phylogeny

Indpendent Contrast	N	r	p
With Eukaryotes			
Genome size vs. Number of Cell Types	139	0.155-0.186	0.029-0.068
ORFs vs. Number of Cell Types	139	0.616-0.641	<0.0001
Without Eukaryotes			
Genome size vs. Number of Cell Types	121	0.225-0.228	0.011-0.013
ORFs vs. Number of Cell Types	121	0.192-0.197	0.030-0.034
Genome size vs. Number of Cell Parts	121	0.278-0.283	0.002
ORFs vs. Number of Cell Parts	121	0.276-0.277	0.002

3.2 Conserved Gene Amino Acid Phylogeny Independent Contrast Analysis

Substantially similar relationships were found using alternative phylogenetic trees derived from conserved protein sequences for 43 species [36], suggesting that these correlations are not an artifact of trees derived from small subunit rRNA sequences. The independent contrast analysis using the Brown et al. (2001) data set showed a significant positive correlation between number of cell types and genome size and between the number of cell types and the number of open reading frames (Table 2). Pruning eukaryotes from the trees and repeating the analysis yielded significant correlations between number of cell types and genome size and between the number of cell types and the number of open reading frames. Continuing to restrict the analysis to prokaryotes and considering the number of types of cell parts gave significant or nearly significant correlations between this quantity and both genome size and the number of open reading frames size.

Table 2. Independent contrast analyses for the conserved gene amino acid phylogeny

Indpendent Contrast	N	r	p
With Eukaryotes			
Genome size vs. Number of Cell Types	43	0.785-0.800	<0.0001
ORFs vs. Number of Cell Types	43	0.899-0.908	<0.0001
Without Eukaryotes			
Genome size vs. Number of Cell Types	39	0.438-0.462	0.004-0.005
ORFs vs. Number of Cell Types	39	0.432-0.459	0.003-0.006
Genome size vs. Number of Cell Parts	39	0.308-0.320	0.047-0.056
ORFs vs. Number of Cell Parts	39	0.345-0.360	0.024-0.031

4 Discussion

For all of the data sets examined here, there are significant positive correlations between genome size or numbers of open reading frames and numbers of cell types and numbers of types of cell parts. These results suggest that the greatest irony about the C-value paradox may very well be that there is no paradox at all and that genome complexity and morphological complexity actually do significantly positively correlate with one another, at least for the organisms with sequenced genomes in this data set. This is not the first time a correspondence between genome size and morphological complexity has been suggested [16, 39, 40], but this is the first time the correspondence is supported by an analysis of independent contrasts that reveals a statistically significant positive correlation. This suggests that organismal morphological complexity may follow some of the same scaling laws that have already been observed in other combinatorial systems [41].

While these results differ from those of most previous studies of the C-value paradox, previous methods for measuring these quantities (such as haploid DNA content, chromosome number, or placement on the scale of the Great Chain of Being [10]) may have been inadequate to detect these correlations. The development of whole genome sequencing and annotation [30, 31] and the creation of new metrics for measuring complexity [19] have permitted this finer-scale understanding of the relationship between morphological complexity and genomic complexity. For those interested in the relationship between genotype and the generation of morphological complexity [42], the detected correlations between numbers of open reading frames and numbers of cell types or types of cell parts suggest that the number of genes present in an organism may have a greater role in permitting, generating, or maintaining morphological complexity than previously anticipated.

A note of caution is warranted in interpreting these results because the selection of genomes to be sequenced has been influenced by genome size, because larger genomes are more costly to sequence. As a result, the tendency has been to select, particularly among eukaryotes, morphologically complex organisms with the smallest possible genome sizes for sequencing. This could predispose data sets containing eukaryotes to reveal positive correlations between genome size and morphological complexity because of issues of taxon sampling. However, the selection of prokaryotes for sequencing, because of their universally much smaller genome sizes, is largely free from this bias, so the analyses of the prokaryote-only data sets included here are probably revealing real positive correlations between measures of genome size and complexity and measures of morphological complexity.

Complete resolution of the C-value paradox will require the consideration of eukaryotic organisms with large genomes and significant amounts of heterochromatin so that a determination can be made concerning whether the relationships reported here also hold at larger genome sizes, something that may not be possible until several organisms with large genomes have been sequenced.

Acknowledgements

The author thanks D. McShea and R. Fehon for the public discussion that inspired this paper, B. Nicklas for his inspirational lectures on genome size, R. Vilgalys for insights into microbial diversity and ultrastructure, J. Brown for the aligned amino acid data set, J. Shapiro and T. Carter for their encouragement, and J. Seiff and B. Marcus for their help and patience. D. McShea, K. Doerner, D. Emlen, T. Evans, A. Harper, K. Hertweck, T. Hughes, T. Powell, B. Polen, and two anonymous reviewers provided helpful comments on the manuscript. Support for this research is from the National Science Foundation and the Commonwealth of Kentucky through an EPSCoR award (EPS-0132295), from the National Institutes of Health and the National Center for Research Resources Grant P20 RR16481, from a scholarship to the Complex Systems Summer School at the Santa Fe Institute, and from a Junior Faculty Scholarship from Western Kentucky University.

References

1. Avery, O.T., C.M. MacLeod, and M. McCarty, Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. J. Exp. Med., 1944. **79**(1): p. 137-158.
2. Watson, J.D. and F.H.C. Crick, A structure for Deoxyribose Nucleic Acid. Nature, 1953. **171**: p. 737-738.
3. Mirsky, A.E. and H. Ris, The deoxyribonucleic acid content of animal cells and its evolutionary significance. J. gen. Physiol., 1951. **34**: p. 451-462.
4. Thomas, C.A., The genetic organization of chromosomes. Annu. Rev. Genet., 1971. **5**: p. 237-256.
5. Cavalier-Smith, T., ed. The evolution of genome size. 1985, John Wiley: New York.
6. Gregory, T.R., Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev., 2001. **76**: p. 65-101.
7. Pagel, M. and R.A. Johnstone, Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. Proc. R. Soc. Lond., 1992. **249**: p. 119-124.

8. Goin, O.B., C.J. Goin, and K. Bachmann, DNA and amphibian life history. Copeia, 1968. **1968**: p. 532-540.
9. Ohno, S., Evolution by gene duplication. 1970, New York: Springer-Verlag.
10. Lovejoy, A.O., The Great Chain of Being. 1936, Cambridge, MA: Harvard University Press. 376.
11. Cavalier-Smith, T., Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci., 1978. **43**: p. 247-278.
12. Cavalier-Smith, T., r- and K-tactics in the evolution of protist developmental systems: cell and genome size, phenotype diversifying selection, and cell cycle patterns. Biosystems, 1980. **12**: p. 43-59.
13. Sessions, S.K. and A. Larson, Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. Evolution, 1987. **41**: p. 1239-1251.
14. Gregory, T.R., Genome size and developmental complexity. Genetica, 2002. **115**: p. 131-146.
15. Gregory, T.R., Macroevolution, hierarchy theory, and the C-value enigma. Paleobiology, 2004. **30**(2): p. 179-202.
16. Doolittle, W.F. and C. Sapienza, Selfish genes, the phenotype paradigm and genome evolution. Nature, 1980. **284**: p. 601-603.
17. Orgel, L.E. and F.H.C. Crick, Selfish DNA: the ultimate parasite. Nature, 1980. **284**: p. 604-607.
18. Nelson, K.E., Paulsen, I.T., Heidelberg, J.F., and Fraser, C.M., Status of genome projects for nonpathogenic bacteria and archaea. Nature Biotechnology, 2000. **18**: p. 1049-1054.
19. McShea, D.W., Functional complexity in organisms: Parts as proxies. Biol. Philos, 2000. **15**(5): p. 641-668.
20. Sneath, P.H.A., Comparative biochemical genetics in bacterial taxonomy, in Taxonomic Biochemistry and Serology, C.A. Leone, Editor. 1964, Ronald Press: New York. p. 565-583.
21. Valentine, J.W., A.G. Collins, and C. Porter Meyer, Morphological complexity increase in metazoans. Paleobiology, 1994. **20**(2): p. 131-142.
22. Carroll, S.B., Chance and necessity: the evolution of morphological complexity and diversity. Nature, 2001. **409**(6823): p. 1102-1109.
23. Bell, G. and A.O. Mooers, Size and complexity among multicellular organisms. Biol. J. Linn. Soc., 1997. **60**: p. 345-363.
24. Bonner, J.T., The evolution of complexity by means of natural selection. 1988, Princeton, NJ: Princeton University Press. 260.
25. Harvey, P.H. and M.D. Pagel, The comparative method in evolutionary biology. 1991, Oxford: Oxford University Press.
26. Felsenstein, J., Phylogenies and the comparative method. Am. Nat., 1985. **125**: p. 1-15.
27. Garland, T., Jr., P.H. Harvey, and I.R. Ives, Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol., 1992. **41**: p. 18-32.
28. Valades, D., Rhetorica Christiana. 1579: Pervsiae, apud Petrumiacobum Petrutium. 10.
29. Fletcher, A., Gender, Sex, and Subordination in England 1500-1800. 1995, New Haven: Yale University Press. 442.
30. CBS Genome Atlas Database. 2003, Center for Biological Sequence Analysis, http://www.cbs.dtu.dk/services/GenomeAtlas/: Lyngby, Denmark.
31. GOLD Genomes OnLine DataBase. 2003, Integrated Genomics, http://igweb.integratedgenomics.com/GOLD/: Chicago, IL.

32. Martins, E.P., COMPARE, version 4.4. Computer programs for the statistical analysis of comparative data. 2001, Department of Biology, Indiana University, Bloomington IN.
33. National Center for Biotechnology Information. 2003, National Library of Medicine, http://www.ncbi.nlm.nih.gov/: Washington, D.C.
34. Jeanmougin, F., et al., Multiple sequence alignment with Clustal X. Trends Biochem. Sci., 1998. **23**: p. 403-405.
35. Swofford, D.L., PAUP*, Phylogenetic analysis using parsimony (*and other methods). 1998, Sinauer Associates: Sunderland, Massachusetts.
36. Brown, J.R., et al., Universal trees based on large combined protein sequence data sets. Nat. Genet., 2001. **28**: p. 281-285.
37. Nelson, K.E., et al., Status of genome projects for nonpathogenic bacteria and archaea. Nature Biotechnology, 2000. **18**(10): p. 1049-1054.
38. Marcus, J.M. and A.R. McCune, Ontogeny and phylogeny in the northern swordtail clade of Xiphophorus. Syst. Biol., 1999. **48**(3): p. 491-522.
39. Rees, H. and R.N. Jones, The origin of the wide species variation in nuclear DNA content. Int. Rev. Cytol., 1972. **32**: p. 53-92.
40. Sparrow, A.H., H.J. Price, and A.G. Underbrink, A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: some evolutionary considerations. Brookhaven Symp. Biol., 1972. **23**: p. 451-494.
41. Changizi, M.A., Universal Scaling Laws for Hierarchical Complexity in Languages, Organisms, Behaviors and other Combinatorial Systems. J. Theor. Biol., 2001. **211**: p. 277-295.
42. Hedges, S.B., et al., A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol. Biol., 2004. **4**: p. 2 doi:10.1186/1471-2148-4-2.