A decade of access to whole-genome sequences has been increasingly revealing about the informational network relating all living organisms. Although at one point there was concern that extensive horizontal gene transfer might hopelessly muddle phylogenies, it has not proved a severe hindrance. The melding of sequence and structural information is being used to great advantage, and the prospect exists that some of the earliest aspects of life on Earth can be reconstructed, including the invention of biosynthetic and metabolic pathways. Still, some fundamental phylogenetic problems remain, including determining the root — if there is one — of the historical relationship between Archaea, Bacteria and Eukarya.

Addresses
Department of Chemistry & Biochemistry, University of California, San Diego, La Jolla, CA 92093-0314, USA
Corresponding author: Doolittle, Russell F (rdoolittle@ucsd.edu)

Introduction
Historically, the primary goal of the study of molecular evolution is to reconstruct past events in a way that explains the present living world. Ultimately, if the evidence has not been overly blurred by time, all trails should lead back to a common ancestral cell type. Over the years, macromolecular sequence information has been applied effectively towards this end, even in the face of major complications resulting from vastly unequal rates of change along different lineages, horizontal transfers of genes and gene clusters, and numerous other distractions. That these efforts have succeeded as well as they have must be regarded as a major triumph.

Although the enterprise has been ongoing for half a century, it’s only during the past decade that whole-genome sequences have been available [1]; the question needs to be asked how this resource has affected the quest. In a word, immensely. Not only are organism connections at all levels being better established, but the full extent of the proliferation of gene families and the protein structures that underlie cellular divergences is also being greatly extended. In this brief review, I attempt to highlight some of the most impressive advances that whole-genome studies have contributed to our views of evolution.

Gene recognition
From its beginning, the whole-genome enterprise depended heavily on the premise that most genes would be readily identified by computer analysis alone. The basis of this hope was that most—if not all—extant genes are descendants from a smaller ancestral population that has been expanded by gene duplication. As such, identifications would be made by comparison with known genes and gene products whose functions had been determined experimentally. Lurking beyond the simple hope that a function would be assigned to every identified gene was an even more optimistic view: that it would eventually be possible to match every putative gene product with a known homologous three-dimensional structure.

It was somewhat disappointing, then, to find that, among the first half-dozen microbial genomes to be completed, almost half of all open reading frames (ORFs) were URFs (the ‘U’ denotes ‘unidentified’) [2]. After more than a century of study by biochemists and microbial geneticists, how could there be so many unrecognizable genes? Were these unrecognizable ORFs the consequence of anomalously accelerated rates of sequence change? Were all these ORFs really true genes? In fact, many of them were quite short and might not even be expressed. But others fell well within the size range of average genes and were found also in other genomes, suggesting they were indeed authentic.

The trend continued as more whole genomes were reported, the fraction of ORFs without identifiable counterparts in sequence databases hardly declining [3,4]. The absolute number grew to the point at which there were 20 000 ORFans (formerly URFs) gathered and cataloged in a single microbial database [5,6,7]. On the bright side, the fraction seems now to be diminishing as searching regimens improve [8,9]. For example, the use of a fold recognition algorithm is making connections that were previously missed when only sequences were being considered [10,11]. Still, it remains mysterious how these ORFans have become so different from their alleged nearest relatives.

Whole-genome trees
The appearance of the first several whole-genome sequences quickly led to attempts to reconstruct
phylogenetic trees based on them. Every possible derivation of the information was put to use, it seemed, and whole-genome trees were generated based on sequence [12–15] and gene content [16,17], as well as on structural attributes (treated separately below). Not surprisingly, there was a degree of incongruity among trees made by different strategies, although combinations of various methods led to consensus trees that seemed reasonable [18,19]. Nonetheless, there were problems and surprises.

Horizontal gene transfers
One complication that arose during attempts to generate whole-genome phylogenies had to do with a substantial number of genes that, when examined by standard phylogenetic methods, were clearly out of order with the parental genome trees. This was particularly true with regard to prokaryotic genomes. The simplest explanation was that the genes had been transferred horizontally. At first glance, the unexpectedly large number of such occurrences seemed to make the prospect of establishing a stable phylogeny unlikely [20]. With time and further reflection, however, it became clear that phylogenetic construction was not going to be seriously impaired, so long as one remained alert to the very real prospect of some genes or gene clusters having been acquired laterally rather than by vertical descent [21–23,24] [25].

In many cases, horizontal gene transfers are extremely interesting on their own [25]. Indeed, many reflect some of the most innovative adaptations in all of biology — including bacterial photosynthesis and nitrogen fixation, both probably having resulted from a series of cellular ‘barters’ [26,27]. Such transfers are not restricted to single genes; genes, operons and much more massive assemblies are commonly exchanged among prokaryotes. Sometimes, as in the case of ‘pathogenicity islands’, they can involve virtual armadas of biological warfare agents [4].

Gene loss
Another phenomenon that became ‘visible’ when every gene in a genome could be counted was the loss of genes along different lineages [28]. It was found, for example, that almost 400 genes were lost along each of the lineages leading from the common ancestor of fission and budding yeasts [29]. In the pre-genome era, most biologists regarded gene loss as a truly calamitous event that could affect the entire interaction network within a cell. That so many genes could be lost in such a relatively short time implied a much more fluid genome than had been anticipated. Significantly, a large fraction of horizontally transferred genes is quickly lost [30]. If the moment of opportunity is not exploited, the event is doomed. In the final analysis, there is good reason to think that gene loss is the principal determinant of gene content [31,32].

Reduced genomes
A variety of whole-genome sequences have been determined for parasitic organisms that have adapted to an existence with severely reduced genomes. These organisms have jettisoned much of their own enzymatic equipment and live off the metabolic resources of their host cells. The phenomenon occurs in all three superkingdoms (Table 1). Examples among Bacteria include familiar organisms such as Mycoplasmas [33], Chlamydias [34], Buchners [35] and Rickettsia [36]. The last named being especially interesting because of a kinship with likely antecedents of mitochondria. Adaptations to parasitic existence are idiosyncratic, different parasite genomes losing different sets of genes on their way to dependence [4]. In this regard, a fascinating situation is afforded by Mycobacterium leprae, a bacterium that is still in the process of decay and that still contains a slew of inactive genes on their way to random oblivion or elimination [37].

A very different situation exists in a unique archaeal parasite, Nanoarchaeum equitans, already firmly established with a minimal genome [38]. N. equitans gains most of its metabolic needs from another archaeon, Igniococcus, to which it is obligately attached [38]. By contrast, Igniococcus can live very well without the parasite. In passing, it is worth noting that — although archaeal endosymbionts (as opposed to parasites) are common in eukaryotic cells, especially in amoebic cells [39] — archaeal pathogens have never been found.

Reduced parasitic genomes are also found among the Eukarya, the first of which to be determined was the microsporidium Encephalitozoon cuniculi, an obligate intracellular pathogen [40]. The apicomplexan Cryptosporidium parvum, although not as small as E. cuniculi (Table 1), is interesting in that even its organelles (plastids and mitochondria) have lost their DNA [41].

Minimal genomes
The initial reports of small bacterial genome sequences led to speculation as to what would constitute a minimal set of genes [42]. Reduced genomes are prisoners of their history, having descended from more complex

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size (Mb)</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanoarchaeum equitans</td>
<td>0.49</td>
<td>536</td>
</tr>
<tr>
<td>Mycoplasma genitalium</td>
<td>0.58</td>
<td>484</td>
</tr>
<tr>
<td>Buchnera aphidicola</td>
<td>0.64</td>
<td>596</td>
</tr>
<tr>
<td>Chlamydia trachomatis</td>
<td>1.04</td>
<td>895</td>
</tr>
<tr>
<td>Rickettsia prowazekii</td>
<td>1.11</td>
<td>835</td>
</tr>
<tr>
<td>Mycobacterium leprae</td>
<td>3.2</td>
<td>1604</td>
</tr>
<tr>
<td>Encephalitozoon cuniculi</td>
<td>2.5</td>
<td>1996</td>
</tr>
<tr>
<td>Cryptosporidium parvum</td>
<td>9.1</td>
<td>3807</td>
</tr>
</tbody>
</table>
circumstances, and the minimal sets of genes they may contain are clearly different from the minimum needed by a free-living organism. In this regard, it is of interest that systematic gene inactivation experiments have shown that a free-living bacterium such as *Bacillus subtilis*, which ordinarily has about 4100 genes, only requires 271 of them to survive [43]. A recent theoretical analysis of what would be the minimal set of genes for a free-living bacterium arrived at the similar number of 206 [44*].

These numbers must be regarded skeptically. They are based on what we know about modern and sophisticated organisms. It's the primitive ones that need to be reckoned. The first step to take backwards in time may be to identify the gene content of the last common ancestor of all current life: the number of genes involved in that hypothetical case usually being thought to be 1000–2000, if only because that is how many genes are found in the smallest extant free-living organisms. The real challenge will be to determine the gene content of the earliest cells, which, at some point, must have been very small indeed.

**Introns, splicing and the origin of eukaryotes**

Quite apart from it being more difficult to identify genes in eukaryotes because of the intronic disruption of coding regions, we might also ask what effect whole-genome studies have had on the long-standing ‘introns early-introns late’ debate. In fact, it seems to have provided ammunition for both sides. That introns are gained and lost by modern eukaryotic genomes at a confounding rate seems indisputable [45,46], the gains being regarded by the ‘introns late’ school as clear support for their side [47]. Contrarily, the fact that losses seem to outnumber gains is used as an argument to support the ‘introns early’ position [48]. It seems to me that ‘early or late’ should not have been made the crux of the argument; rather, the focus should be on those original claims that extant exons are vestigial remnants of the earliest proteins [49]. A full appreciation of the history of protein structures requires that these matters be settled more fully [50].

**Refining relationships**

Whole-genome projects are definitely improving the overall quality of phylogenies, chipping away, for example, at the thorny problem of how the major bacterial phyla are related at the deepest levels [18]. Archaeal phylogenies need similar study, as evidenced by ribosomal trees usually not being congruent with whole-genome trees made on the basis of non-ribosomal attributes. That such contradictions can be overcome when sufficient data are brought to bear is shown by the recent successful clustering of very distantly related amoebas into a clade that conforms perfectly to classical biology [51*,52].

**The interactome**

Over and beyond tracing the history of organisms and their proteins, there has been significant progress made in determining which gene products interact with each other and how the general outlines of metabolism evolved. Previously, interactions between macromolecules had to be determined experimentally, the yeast two-hybrid system having become the strategy of choice for finding interacting macromolecules [53]. The initial *in silico* tactic was simply to look for genes that were adjacent or very near to each other more often than would be expected [54,55]. Methods have improved to the point at which they now rival experimental approaches [56]. The interaction of its proteins and other macromolecules is the backbone of the cell’s machinery, upon which its survival depends.

**Melding sequence and structural attributes**

From the start, the whole-genome project was the beneficiary of remarkable advances in structural biology. During the past decade, the number of three-dimensional structures in the Protein Data Bank (PDB) [57] has swelled to more than 20 000 entries. Concurrently, the SCOP (Structural Classification of Proteins) database [58,59] has been parsing the PDB structures into their constituent domains — more than 50 000 in version 1.65 — all sorted hierarchically by structural type. The process begins by assigning each domain to one of the six major groups (all-α, all-β, α/β, α+β, membrane associated and small). In version 1.65, these have been further grouped into 800 folds, followed by a further subdivision into 1294 superfamilies. The fold level is defined as comprising domains with similar arrangements of secondary structure elements but which may not necessarily reflect common ancestry. The superfamily level of folds is defined as those folds for which there is good evidence of common ancestry, even if sequence similarity is not obvious. The family level is reserved for those members for which sequence similarity reflecting common ancestry is obvious.

The bridge to whole-genome biology is currently provided by the SUPERFAMILY database [60,61]. Thus, every ORF from every whole genome is searched against the SCOP holdings in an effort to match it with a known structure. Most current search procedures rely on sensitive hidden Markov models [62], the effectiveness of which is attested to by more than 60% of ORFs from 174 completed genomes having been assigned to domain superfamilies [61].

**Protein folds and whole genomes**

It didn’t take long for structural biologists to tally up the putative domain structures in the newly determined whole genomes [63–69,70*,71]. These studies have taken several directions, but all managed to count the different folds in the various superkingdoms and to show
how phylogenies could be rendered from them. These studies also provided data about the relative abundances of different kinds of protein domains overall, there being general agreement that α/β domains are the most common structural element. There was much agreement, also, that the distribution of structural features is in general accord with the tripartite nature of the ‘tree of life’, Archaea, Bacteria and Eukarya having distinguishable fold contents [63–69,70**,**71].

The tripartite tree of life

Nonetheless, the conundrum of how that triumvirate evolved remains. Broadly speaking, there are three general schools of thought on the matter. First, there are those who feel that a divergence leading, on the one hand, to a cytoskeleton-containing pre-eukaryote, on the other, to prokaryotes (including the ancestors of both Archaea and Bacteria) occurred very early. Subsequently, a series of phagocytic events (in which pre-eukaryotes engulfed prokaryotes) gave rise to modern eukaryotes [72–75]. Another group feels that the invention of the Eukarya was the result of a chimeric merging of a member of the Bacteria with an archaeon, the latter assuming the form of the nucleus in the new entity [76,77]. A third interpretation is that all three lines originated at a time when the totality of life on Earth was a simple community of heterogeneous cells that shared only the ability to synthesize proteins with a ribosomal machine and that freely exchanged genetic material [78–80]. Will whole-genome biology be able to answer this question? Let us be hopeful.

Pushing backwards in time

Disputes about the three major domains of life aside, progress is being made concerning events that must have occurred well before the last common ancestor, especially with regard to protein structures. Aravind et al. [81] have reported a convincing analysis of more than a dozen kinds of nucleotide-binding domains that occur in all living organisms (‘Rossmannoid domains’). They constructed a phylogenetic tree depicting the evolution of the various types from a common structure that logically pre-dates the last common ancestor.

In an even bolder maneuver, Caetano-Anolles and Caetano-Anolles [70**] constructed phylogenetic trees of all known protein folds. Using strict cladistic principles and straightforward measurements of fold usage as a fundamental character for a hierarchical reconstruction, they showed how α/β proteins originated first and have given rise to the other main protein classes. Moreover, they were able to depict logical scenarios for the evolution of known folds within each of the main classes. Although the underlying premise of this approach — that “redundancy is a favored evolutionary outcome” — might be challenged, these are enormously interesting studies that deserve much scrutiny and discussion. It can’t be a coincidence that α/β domains are the most predominant type of structure in contemporary proteins and that the glycolytic pathway is composed of proteins that are almost exclusively made from such domains.

Conclusions

The first decade of whole-genome biology has been exciting; it has taught us a great deal about how genomes evolve. But there is much more to come and the next decade should teach us a lot more. There is definite promise that the clarity of seeing backwards in time will improve. I feel confident that the notion of the last common ancestor will be revalidated and — I hope — the controversy over the root of the three-superkingdom triad finally settled. One approach might be to construct trees based on ‘interactome’ comparisons. For example, actin is well known to have a multitude of interactants. It should be possible to reconstruct the history of how these various associations accumulated, during the course of time, in parallel with eukaryotic attributes such as phagocytosis. In the end, we may be able to glimpse behind the curtain that separates us from life before the last common ancestor, perhaps even into an ancient RNA-protein world.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest


25. The perils of disconnecting horizontal gene transfers in constructing phylogenies are emphatically put forth.


82. Among other things, a scenario is depicted suggesting the early diversification of the Rossmann fold.