Angela Torres
Bioc258 Final Paper
12/7/13

Bioinformatics and Insights into the Origin of Life

Hundreds of years ago, when Darwin observed the theory of evolution at work in the Galapagos, he could not have imagined that a complex mechanism of information transfer was the basis of his theory. He postulated that environmental influences fundamentally affected the biology of an organism, foreshadowing the networks of gene-environment interactions that we observe today.

There are many instances of evolution that are present in our daily lives. Perhaps the strongest examples of this mechanism at work are seen in the evolution of microbial and viral pathogens-influenza, HIV/AIDS, and the rise of antibiotic resistance. Indeed, instances of bacterial transfer of information and subsequent changes and adaptations are very clear on a microscopic level. However, the conclusive evidence for the most audacious claim of evolutionary theory, namely that all organisms are descendents of a common ancestor, remain elusive as first-hand observation would require breaking the laws of physics to travel through time. In addition, the origins of our genetic code remain a mystery.

However, the discovery of DNA structure, its RNA messenger, as well as the striking genomic similarities across animal species and families have allowed us a snapshot into our origins. This paper will explore the revelations of bioinformatics to the study of our origins, and suggest future steps for the future of evolutionary bioinformatics.

Angela Torres
Bioc258 Final Paper
12/7/13

Comparative Genomics: Reconstructing evolutionary history

Before the current age of available genomic data, charts of evolutionary relatedness, or phylogenies were constructed by tracing a single protein or gene family for which sequence data was available. Fitch and Margoliash describe a method in which they identify the number of mutations required for a particular amino acid to another amino acid. They then used these calculations to identify mutational distances between the proteins of different organisms. The lower the mutational distances, the more related the organism. In their paper, they describe the limitations of constructing these trees, namely that sequences for many other proteins were not widely available. (1)

Today, the readily accessible genomes of most organisms allow for the construction of phylogenies based on many different bioinformatics methods. Phylogenies may be alignment-free, using a distance-based clustering method or a shared information computer program that finds the shortest distance to get from genome A to genome B. Trees can also be constructed using shared gene content, genome order, or by following a particular gene family. (2) Many of these methods use the publicly available data sets from NCBI and use BLAST searches to determine molecular and protein similarities to identify orthologs and paralogs. In addition, when extra proteins or genes are needed for closer approximations of divergences or relatedness, next-generation sequencing techniques provide the required information.(3) Combining these phylogenetic histories with the fossil record, molecular clocks, and rates of mutation of the genome and proteosome allow evolutionary biologists to approximate the time of divergence of

species.(3) A web interface, “TimeTree” has aggregated this data to produce a massive phylogeny of the animal kingdom, tracing back to a theoretical Last Universal Common Ancestor (LUCA). (4)

Each of these methods has its downfalls, though most produce very similar phylogenies. The alignment-free method does not take orthologs or sequence homology into consideration, and yet the results are accurate. When analyzing phylogenies based on gene content, one must take into account the size of the genome, as gene content can be overstated if comparing between two organisms with large genomes in relation to more closely related, small-genomed species. In addition, when relying solely on sequence alignments, the distinction between homologs and orthologs is once again lost. (2) Despite these challenges, the methods seem to provide support for phylogenies based on morphological observations, in accordance with what has been observed for decades. However, the search remains for the famed LUCA.

In search of the Last Universal Common Ancestor

The field of comparative genomics has been key in the search for this ancestor. One requirement of LUCA is that it contained the minimal set of genes required for function. However, the concept of a minimal set depends on many factors besides the intracellular machinery. Koonin makes the point that a cell growing on media missing all of the essential amino acids except for cysteine will have a different minimal gene requirement than cell medium that only lacks lysine.(5) These genes are derived using many techniques, including experimental knock-out approaches and comparative genomic approaches. Using computational approaches to

the find the minimal gene set is based on the orthology of proteins in different species. Typically, orthologs are deemed as such because they share common ancestry, and it is often implied that they retain their function. (6) Therefore, tracing orthologous genes down the evolutionary path should provide the minimal set required for cellular function. In addition, the sequencing of very small bacterial genomes are also starting points for finding these essential protein functions, especially since these functions must be retained in a very small instructional set.(7)

One problem that arises from the minimal gene set concept is that even in highly related species, different, non-orthologous proteins may occupy the same functional niche. Thus, many minimal gene sets can be made that would carry out the appropriate functions. Instead, Koonin suggests that the ancestor must contain a set of gene functions: DNA metabolism, RNA metabolism, protein processing and folding, cellular processes (such as cell division), energy metabolism, and other genes whose functions are poorly characterized.(see Figure 1).(5,7) We can further encapsulate these functions by the processes of metabolic homeostasis, reproduction and evolution.(8) Future experimental approaches might be to construct bacteria with knockouts that leave only the minimal gene set. However, due to horizontal gene transfer and non-orthologous gene displacement disrupting the construction of phylogenies, this gene set may be difficult to track.(5)

Tracing the problem back further: DNA origins

So there we have it. Using our computational methods leads us to the creation of a morphologically accurate tree of life, complete with molecular dates. We can even begin to

imagine the genome of our Last Common Universal Ancestor. However, these trees simply take us back to a much larger problem. How did the information that we now use so freely to construct these models come to be in the first place? As a point of comparison, we can take a look at the history of the modern computer. We can create a phylogeny based on the advancements and additions that have been made up to this point. However, can we say the same for the language that drives processes and programs on a computer? This is a massive point of inquiry, and several theories have been put forth for explaining the phenomenon of the evolution of the DNA-RNA information system.

When analyzing the origin of life problem, it is unlikely that the central dogma of genetic transfer was at play when the system of information was developed. The familiar paradigm is that DNA is copied into RNA, processed into messenger RNA, and then translated into protein by the translational machinery. However, without protein, DNA cannot be replicated; without RNA, protein cannot be translated; and without DNA mRNA can't be transcribed. This leaves us with an old *chicken or the egg* question.(9)

A way to avoid this problem, and therefore bypass the necessity and statistical improbability of both protein and DNA evolving together, an "RNA-world" hypothesis has been proposed. Many experiments point to RNA as the original genetic molecule, as it can perform enzymatic functions necessary for metabolism. In addition, it is present in the translational machinery of ribosomes, leading researchers to postulate that the RNA-protein complex that makes up the ribosome might be an early RNA polymerase for RNA replication.(9)

Still, qualms remain regarding the RNA-first hypothesis. First of all, RNA is much more unstable than DNA. In addition, though progress on the search for RNA polymerases has been made, and novel RNA catalysis functions have been discovered through *in-vitro* evolution, (9,10). Trevors argues that this type of evolution is still engineer-driven, and therefore invalid in the discussion of the RNA-world hypothesis. (11)

Despite the persistence of this theory, the leap from RNA to DNA is still an open chasm. How do we get DNA as the genetic molecule? A look into the history of life promotes more questions than answers: How did the three-nucleotide code develop? Can the principles of thermodynamics and chaos and complexity explain the emergence of information systems? (12)

Perhaps these questions cannot be answered by bioinformatics alone, but Mac Dónaill of Trinity College in Dublin suggested that perhaps the 4-letter DNA code is the optimal "choice" that nature had. He asserts that DNA is an information system; therefore informatics considerations must be taken into account in the evolution of the system, in addition to the physical and chemical constraints on the molecule. By translating the 4-letter genetic code into a 4-bit binary system based on (a) the choice between being a purine or a pyrimidine (b) being a hydrogen donor or hydrogen acceptor. In the end, the code for each nucleotide has even parity, which provides an error-prevention mechanism. Being bonded to certain combinations of the 4-bit system are highly unlikely, as there is one natural bond and others would result in weak interactions. (13) Interestingly, the principles of computer science may apply in the study of the evolution of DNA.

Angela Torres
Bioc258 Final Paper
12/7/13

Bioinformatics can provide many clues; homologs, related sequences, phylogenetic trees; into our origins. Indeed, the origin of life question may still remain outside of the realm of this field. However, the underlying principles of bioinformatics may provide more clues as to our internal, literal bioinformatics code. Just as an engineer produces algorithms that are crucial for a program to be carried out, our origins may be based on complex algorithms waiting to be discovered.

Works Cited

1. Fitch WM, Margoliash E. Construction of Phylogenetic Trees. Science. 1967 Jan 20;155(3760):279–84.

2. Snel B, Huynen MA, Dutilh BE. Genome trees and the nature of genome evolution. Annu Rev Microbiol. 2005;59:191–209.

3. Hedges S, Blair J, Venturi M, Shoe J. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol. 2004;4(1):2.

4. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006 Oct 4;22(23):2971–2.

5. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 2003 Nov;1(2):127–36.

6. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38.

7. Gil R, Silva FJ, Pereto J, Moya A. Determination of the Core of a Minimal Bacterial Gene Set. Microbiol Mol Biol Rev. 2004 Sep 7;68(3):518–37.

8. Luisi PL, Oberholzer T, Lazcano A. The Notion of a DNA Minimal Cell: A General Discourse and Some Guidelines for an Experimental Approach. Helv Chim Acta. 2002 Jun;85(6):1759–77.

9. Penny D. An Interpretive Review of the Origin of Life Research. Biol Philos. 2005 Sep;20(4):633–71.

10. Orgel LE. The origin of life—a review of facts and speculations. Trends Biochem Sci. 1998 Dec;23(12):491–5.

11. Trevors J, Abel D. Chance and necessity do not explain the origin of life. Cell Biol Int. 2004 Nov;28(11):729–39.

12. Abel DL. The Capabilities of Chaos and Complexity. Int J Mol Sci. 2009 Jan 9;10(1):247–91.

13. D??naill DAM. A parity code interpretation of nucleotide alphabet compositionElectronic supplementary information (ESI) available: expanded background to information and error-coding theory; computational details. See http://www.rsc.org/suppdata/cc/b2/b205631c/. Chem Commun. 2002 Sep 11;(18):2062–3.