Lining Up to Avoid Bias

Antonis Rokas

D

arwin relied on fossils, morphology, and geographical distribution to glean important clues about the history of life. Today, natural historians can study organisms’ history of change and adaptation by probing the DNA record. Whether to elucidate evolutionary relationships of genes and species or spot the amino acid changes driven by selection, we need to be able to generate accurate alignments of DNA sequences. On page 473 of this issue, Wong et al. (1) provide some important caveats on how this can go awry and how to avoid alignment bias.

Sequence alignment is how we identify the similarities in a set of sequences derived from a common ancestor. For very similar genes or proteins we can do this manually, but the alignment of more divergent sequences that have been peppered not only with substitutions, but also with insertions and deletions (collectively known as indels), is particularly challenging. The uncertainty in identifying the precise number and length of indels as well as their exact placement (2) creates a multitude of potential ways of aligning distantly related sequences, a computational problem that has given rise to a cadre of alignment-generation programs, each based on a different optimization algorithm (3).

For years, the standard protocol has been to pick a favorite algorithm to optimize the alignment it generates. This approach is fast and easy, but it is like being forced to always settle on vanilla ice cream for dessert; doing so can taint one’s opinion about ice cream. Similarly, sticking to the use of a single alignment from a single algorithm can bias the estimation of phylogenetic trees and amino acid changes driven by selection revealed that all alignment methods contributed substantially to the sensitivity of evolutionary analyses to the choice of method. A case in point is the gene YPL077C, whose seven alignments produced six different phylogenies. Furthermore, a Bayesian measurement of alignment variability (see the figure) (5) was significantly correlated with the variability in phylogenetic inference. Consequently, these results argue that the observed uncertainty in inference is likely explained by the fact that some genes were much more difficult to align than others and thus challenging to all alignment methods.

Should we all revisit our evolutionary studies because we haven’t properly accounted for uncertainty in alignment? Some of the disagreements observed across the molecular evolution and comparative genomics literature may be the result of alignment differences, but there is good news too. To begin with, several popular markers for evolutionary analyses became popular in the first place because they are easy to align. These genes are unlikely to exhibit variation in alignment and hence unlikely to be affected by the results of Wong et al., although this is not always so (4).

Furthermore, the authors show that when the seven alignment treatments per gene result in two or more phylogenetic trees, the resulting trees rarely receive high “bootstrap” values (1), a popular index that uses resampling with replacement to measure robustness in infer-

Alignment’s true colors. Alignment uncertainty for a gene fragment from four of the Saccharomyces yeasts species used in the study by Wong et al. Areas of low uncertainty are shown in red (or dark gray for indels), whereas areas of high uncertainty are shown tending toward violet (or light gray). The plot was generated by the program Bali-Phy (5).

Alignment uncertainty for a gene fragment from four of the Saccharomyces yeasts species used in the study by Wong et al. Areas of low uncertainty are shown in red (or dark gray for indels), whereas areas of high uncertainty are shown tending toward violet (or light gray). The plot was generated by the program Bali-Phy (5).
ty, or otherwise, about them. But there’s a catch: The computational demands of these programs are prohibitive.

As in any scientific field, molecular evolution has a long tradition of dramatic transformation. The development of a powerful computational and statistical arsenal to account for the uncertainty stemming from sequence alignments is heralding the first paradigm shift in the era of genome-scale analysis.

References

10.1126/science.1153156

SYSTEMS BIOLOGY

Enlightening Rhythms

Ovidiu Lipan

How yeast systematically respond to environmental change emerges from blending engineering, mathematical, and experimental analyses.

We live in a sea of vibrations, detecting them through our senses and forming impressions of our surroundings by decoding information encrypted in these fluctuations. Such periodic phenomena range from circadian oscillations in living cells (1) to acoustic oscillations in the primordial universe (2). Passively observing periodic phenomena is scientifically rewarding, but actively using periodic stimuli to observe the hidden wonders of nature is even more so. On page 482 of this issue, Mettetal (3) report using oscillatory stimuli to decipher how an organism—the yeast Saccharomyces cerevisiae—responds to environmental changes. By constructing a predictive mathematical model for specific signaling pathways (4), they show that oscillatory stimuli can be used to study how networks of proteins and genes are engaged by a living system to control physiological behavior.

Many scientific studies hinged on creating oscillations to study natural systems. The idea of electromagnetic waves was implicit in James Maxwell’s theory, but it was Heinrich Hertz’s electric oscillators that created and measured their properties, thus confirming light waves as electromagnetic radiation, the most striking victory of 19th-century experimental physics (5).

Although the use of oscillatory stimuli to study how networks of proteins and genes regulate gene expression is theoretically valuable (6), implementation of this procedure is not obvious because the possibilities for constructing genetic oscillators are limited, at present. It takes ingenuity to find a molecular pathway that responds to an oscillatory signal, much less an experimental procedure to create these oscillations. Furthermore, these oscillations must produce detectable responses. Mettetal et al. fulfill these constraints by studying a signaling pathway in yeast that responds to changes in environmental osmolality. Glycerol is the main yeast osmolyte and its concentration is controlled in part by the high-osmolality glycerol (HOG) signaling pathway that involves the enzyme Hog1. By adjusting the export rate of glycerol through the cell membrane, yeast maintain osmotic balance.

Mettetal et al. studied three negative-feedback loops of the HOG pathway. One loop controls glycerol concentration through the membrane protein Fps1, and depends on the amount of active Hog1 in the nucleus. A second loop also involves Fps1, but is controlled by osmotic pressure across the cell’s membrane and the concentration of intracellular glycerol. A third loop is Hog1 dependent and acts on glycerol concentration by increasing the expression of genes encoding the glycerol-producing proteins Gpd1 and Gpp2.

Which of the three negative-feedback loops dominate the dynamics of this osmo-adaptation system? Can system identification methods, such as those used by robotics engineers, describe the signaling dynamics of the dominant negative-feedback loops? To apply systems engineering methods, an input signal must be created and an output response signal must be recorded. Mettetal et al. varied the concentration of sodium chloride in the cell media, thus exposing cells to square-wave pulses (alternating between two values for an equal amount of time) of osmotic pressure. The output response recorded was the ratio between nuclear-localized, active Hog1 and Hog1 within the entire cell.

For designing complex systems, a success is claimed if the output response of a system to an external input signal can be mathematically predicted. Viewed only in terms of its input and output characteristics, the osmo-adaptation pathway loses its inner biological structure and becomes what is referred to as a “black box.” The authors developed a black-box mathematical model for the osmo-adaptation pathway. They estimated parameters of the model from measurements using square-wave pulses of variable frequencies, then validated the predictive power of the model using a step input—one that switches on at a definite time and remains on indefinitely—of sodium chloride. Because gene regulatory networks contain many unknown molecular components, a black-box mathematical model is the best achievable solution in various situations.

The ultimate goal, however, is a mathematical model for a white box, in which all the molecular components and their interactions are known. The road toward this goal is paved with intermediate gray-box models containing some biological inner structures. Toward this end, Mettetal et al. transform the black-box mathematical model into a gray one that successfully incorporates the first two of the three osmo-adaptation feedback loops described. In doing so, they discovered that the dynamics of the osmo-adaptation response are dominated by the fast-acting Hog1-dependent negative feedback loop that does not require a change in gene expression.

The hope is to include other molecular components and feedback loops into a