Molecular messages

Functional information

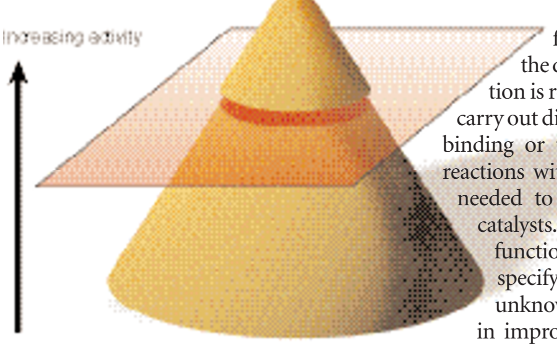
A quantitative means of comparing the functional abilities of different biopolymers would allow us to dissect out differences and to discern their origins.

Jack W. Szostak

In this age of genome sequencing, the idea that biopolymer sequences are a type of molecularly coded information is well established. We are all familiar with the idea that it is the sequence of the nucleotides or amino acids that make up DNA, RNA or protein molecules that determine their structure and function. But the recent deluge of phylogenetic sequence data provides thousands of examples of related but different sequences encoding essentially identical structures and functions. More radical are the accumulating examples of both RNA and protein molecules with entirely different structures but similar biochemical functions (for example, various structurally distinct protease enzymes have been identified). Such examples raise important questions about the nature of the information content of biological sequences. How best can we define and quantify the information content of biopolymer sequences?

The information content of biopolymers is usually thought of in terms of the amount of information required to specify a unique sequence or structure. This viewpoint derives from classical information theory, which does not consider the meaning of a message, defining the information content of a string of symbols as simply that required to specify, store or transmit the string. Thus, the unannotated human genome sequence can be encoded in a 750-megabyte file, but this could be greatly reduced in size by the application of standard data-compression techniques to account for internal repetitions.

Approaches such as algorithmic complexity further define the amount of information needed to specify sequences with internal order or structure, but fail to account for the redundancy inherent in the fact that many related sequences are structurally and functionally equivalent. This objection is dealt with by physical complexity, a rigorously defined measure of the information content of such degenerate sequences, which is based on functional criteria and is measured by comparing alignable sequences that encode functionally equivalent structures. But different molecular structures may be functionally equivalent. A new measure of information — functional information — is required to account for all possible sequences that could potentially carry out an equivalent biochemical function, independent of the structure or mechanism used.



Increasing activity implies fewer sequences and greater functional-information content.

By analogy with classical information, functional information is simply $-\log_2$ of the probability that a random sequence will encode a molecule with greater than any given degree of function. For RNA sequences of length n, that fraction could vary from 4^{-n} if only a single sequence is active, to 1 if all sequences are active. The corresponding functional-information content would vary from $2n$ (the amount needed to specify a given random RNA sequence) to 0 bits. As an example, the probability that a random RNA sequence of 70 nucleotides will bind ATP with micromolar affinity has been experimentally determined to be about 10^{-11}. This corresponds to a functional-information content of about 37 bits, compared with 140 bits to specify a unique 70-mer sequence. If there are multiple sequences with a given activity, then the corresponding functional information will always be less than the amount of information required to specify any particular sequence. It is important to note that functional information is not a property of any one molecule, but of the ensemble of all possible sequences, ranked by activity.

Imagine a pile of DNA, RNA or protein molecules of all possible sequences, sorted by activity with the most active at the top. A horizontal plane through the pile indicates a given level of activity; as this rises, fewer sequences remain above it. The functional information required to specify that activity is $-\log_2$ of the fraction of sequences above the plane. Expressing this fraction in terms of information provides a straightforward, quantitative measure of the difficulty of a task. More information is required to specify molecules that carry out difficult tasks, such as high-affinity binding or the rapid catalysis of chemical reactions with high energy barriers, than is needed to specify weak binders or slow catalysts. But precisely how much more functional information is required to specify a given increase in activity is unknown. If the mechanisms involved in improving activity are similar over a wide range of activities, then power-law behaviour might be expected. Alternatively, if it becomes progressively harder to improve activity as activity increases, then exponential behaviour may be seen. An interesting question is whether the relationship between functional information and activity will be similar in many different systems, suggesting that common principles are at work, or whether each case will be unique.

The challenge in determining experimentally the relationship between functional information and activity is the extreme rarity of functional sequences in populations of random sequences (typically 10^{-10} to 10^{-15} for aptamers and ribozymes isolated from random RNA pools). *In vitro* selection and amplification allow the isolation of rare functional sequences from a large initial pool of random sequences. Unfortunately, the original distribution of functional molecules can be obscured by biases in replication and selection efficiency that accumulate over cycles of enrichment. A radically different approach would be to apply the new single-molecule fluorescence methods to the direct analysis of large sets of random sequences. Such experiments might ultimately allow us to understand why proteins have taken over so much of biochemical function from RNA, and they might also serve to guide and interpret the results of experiments in which new nucleotides or amino acids are used to expand the genetic code as we search for molecules even better than those supplied by nature. ■

Jack W. Szostak is in the Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114-2696, USA.

FURTHER READING

Hamming, R. W. *Coding and Information Theory* (Prentice-Hall, Englewood Cliffs, New Jersey, 1987).

Zurek, W. H. in *Studies in the Sciences of Complexity* Vol. 8 (ed. Pines, D.) (Addison-Wesley, Reading, Massachusetts, 1991).

Adami, C. & Cerf, N. J. *Physica D* **137**, 62–69 (2000).

Wilson, D. S. & Szostak, J. W. *Annu. Rev. Biochem.* **68**, 611–648 (1999).

Erratum

In Antonio Damasio’s Concepts essay on “Mental self” (*Nature* **423**, 227; 2003), the name of Gus Nossal was misspelt.