# Information hourglass and information fountain

*Change L. Tan*

The central dogma of molecular biology formulated by Francis Crick,[1,2] and modified and popularized by James Watson[3,4] was regarded as the only exception to the "'ubiquitous exception' rule" of biology in which "the only actual rule is that there are no rules, i.e. exceptions can be found to every 'fundamental' principle if one looks hard enough",[5] and as *the* theoretical framework for molecular biology.[6] In Crick's words:

> "The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid."[7]

However, we demonstrated recently that the central dogma must be revised in both the concept of information and its transfer.[8] Here we will briefly summarize why and how we want to revise the central dogma and provide a new idea on biological information transfer.

## Concept of biological information

Regarding the concept of information, there are different kinds and different levels of biological information encoded in genomic DNA. Protein coding is only part of its encoded information (for a detailed analysis, see reference 8). We proposed that cellular information can be divided into two kinds, sequence information and episequence information. 'Sequence information' refers to the nucleotide sequence of DNA and RNA, as well as the amino acid sequence of proteins. 'Episequence information' refers to relevant cellular context information, including DNA, RNAs, proteins, metabolites, and other molecules inside a cell at the specific moment under consideration; plus their chemical modifications, localization, structures, concentration, and interactions. Crick's central dogma deals with only the sequence information, and only its protein-coding part. Note that the episequence information includes all molecular machines inside the cell, including the cellular machineries for DNA replication, transcription, translation, and metabolism.

The two kinds of information not only coexist inside cells but also work together to sustain life. First, the episequence information interprets or decodes the sequence information. Second, correct interpretation occurs only when the coding and the decoding systems match with each other. Together, they determine whether the sequence information is meaningful, and, if so, what it means and whether and how it should be expressed (e.g., whether a segment of DNA encodes a gene, whether a gene is transcribed or translated). Third, sequence information affects episequence information. For example, the structure of a protein can be greatly affected by its sequence. A specific DNA or protein modification may only occur to a nucleotide (for DNA) or an amino acid (for protein) when it is localized within a specific sequence motif (or context). That is, episequence changes may rely on specific sequence patterns.

Last, opposite to what one may expect based on Crick's central dogma and our familiar genetic codon table, there is no fixed one-to-one 'residue-by-residue' biological sequential information. Instead, the meaning of a specific DNA segment depends on the system (i.e., the information transfer mechanism of a specific organism), the sequence context,

and the cellular context (i.e., the episequence of the cell). For example, the same RNA input sequence could be decoded by bacterial or eukaryotic translation machinery, generating entirely unrelated peptide sequences, if any.[8] Furthermore, even in the same organism, in the same cell, in the same gene, the same nucleotide sequence can have different meanings depending on its sequence and episequence contexts. For instance, three consecutive AUG nucleotides, when located within a protein-coding region, can be a translation starting site, a methionine in the middle of a protein, parts of codons for other amino acids, even parts of stop codons (…A *UGA*, *UGA* UG…, *UAA* UG… stop codons italicized, and AUG underlined), depending on its location relative to the protein-coding region and its reading frame. The same AUG nucleotide series will not be a codon, or part of any codons, if it is located outside the protein-coding regions. Note that whether a specific AUG is recognized as being located within a protein-coding region varies with the transcription and translation mechanism of the organism (see figure 3 of reference 8).

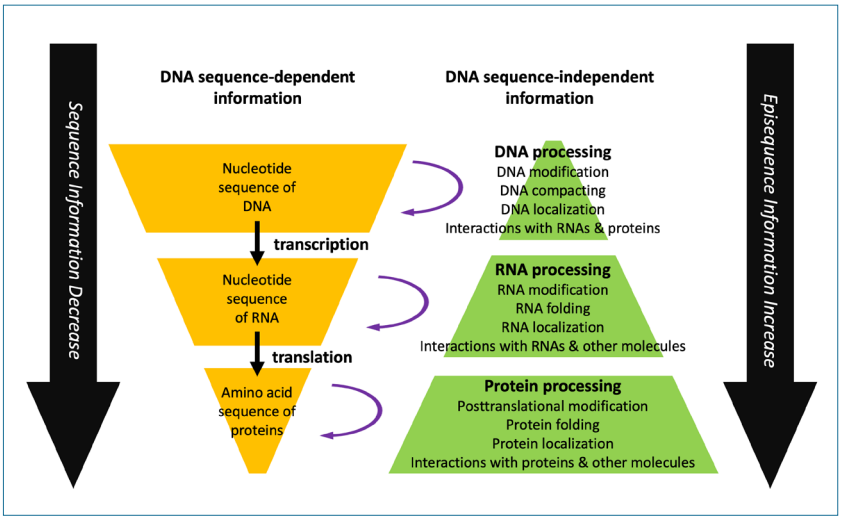


**Figure 1.** The information gain and loss funnels. Left: Loss of sequence information from DNA to RNA to proteins during transcription and translation. Right: Gain of episequence information from DNA to RNA to proteins during or after transcription and translation. (After figure 10 in ref. 8.)

## Characteristics of biological information transfer

### Prohibited transfer

Crick stated, in the central dogma, that sequence information "cannot be transferred from protein to either protein or nucleic acid." This has been referred to as the great biological exclusion principle[9] and as something trivial with "no practical significance to science".[10] Koonin argued that this exclusion is due to "the transition from the digital information carriers, nucleic acids, to analog information carriers, proteins, which involves irreversible suppression of the digital information."[9] On the other hand, Hubert Yockey argued that it is due to the genetic code belonging to a class of codes known as non-isomorphic codes.[11] We argue, instead, that no information, including the sequence information from nuclei acid to proteins, can be transferred unless the coding system and the decoding system match with each other, and under the correct internal and external environment of the cell.[8] The point is that sequences can be used to communicate many unrelated intentions. Each of these codes relies on distinct decoding hardware.

### Information gain and loss

In the light of our new biological information concept, we described a novel way of interpreting information gain and loss.[8] Specifically, we described two information funnels (figure 1). On the one hand (figure 1, left), sequence information is lost when a gene is transcribed and translated. This is because not all regions of genomic DNA are transcribed, and not all regions of RNA are translated (not to mention that there are many RNAs that do not encode for any proteins and are never translated). Consequently, one cannot deduce the whole genome sequence based on the sequence of RNAs or proteins present in the cells containing that genome. On the other hand (figure 1, right), episequence information is incorporated in the same process to ensure correct kinds and levels of gene products are produced according to the type and status of the cell and its internal and external environment. For example, an RNA may only be spliced in the presence of a particular splicing factor.[12] From the same mRNA encoding the bacterial release factor 2 (RF2), either a functional full length or a non-functional truncated RF2 will be generated, depending on the concentration of RF2.[13]

### An information transfer hourglass

In that article, we also introduced the idea of an information transfer hourglass (figure 2).[8] On the top is the determination of whether a DNA molecule is replicable, whether a segment of DNA encodes any gene, whether a gene is protein-coding (i.e., protein as its end product) or not (i.e., RNA as its end product), and whether the gene should be expressed (i.e., transcribed or translated). On the

bottom are specific gene products, including specific RNAs and proteins. Both ends vary with organisms, tissues, cell-types, cell status, and environmental conditions. In contrast, the middle appears to be the same for all currently known organisms. This includes compositional monomers of nucleic acids or proteins, the chemical linkages among the monomers in their corresponding polymers, and the chemical reactions involved in monomer activation and polymerization.

Note that our information transfer hourglass is not a typical hourglass. The two ends of our hourglass are interdependent and connected not only via the narrow hourglass neck but also lines like those used to describe the magnetic field. The processes and decision points on the top influence and impact the resulting genes and gene products on the bottom (information flows through the middle part of the hourglass, like a regular hourglass), while the latter influence and impact the former (via the peripheral 'magnetic field lines', which are absent in a regular hourglass).

Note also that our 'magnetic' lines are directional; they cannot go backward directly because there is no reverse gene translation. In other words, when our hourglass is flipped, the 'time-sand' (the amino acid sequence information of a protein) in the hourglass cannot flow backwards (cannot be back-translated into the nucleotide sequences of its coding DNA or RNA by cells).

Further, the constituents at one end must match—biochemically, physically, and in terms of information transfer capability—those at the other, forming a single integrated, coordinated, and coherent system. That is, the information coding and decoding systems must match with each other for appropriate biological information transfer; thus, the

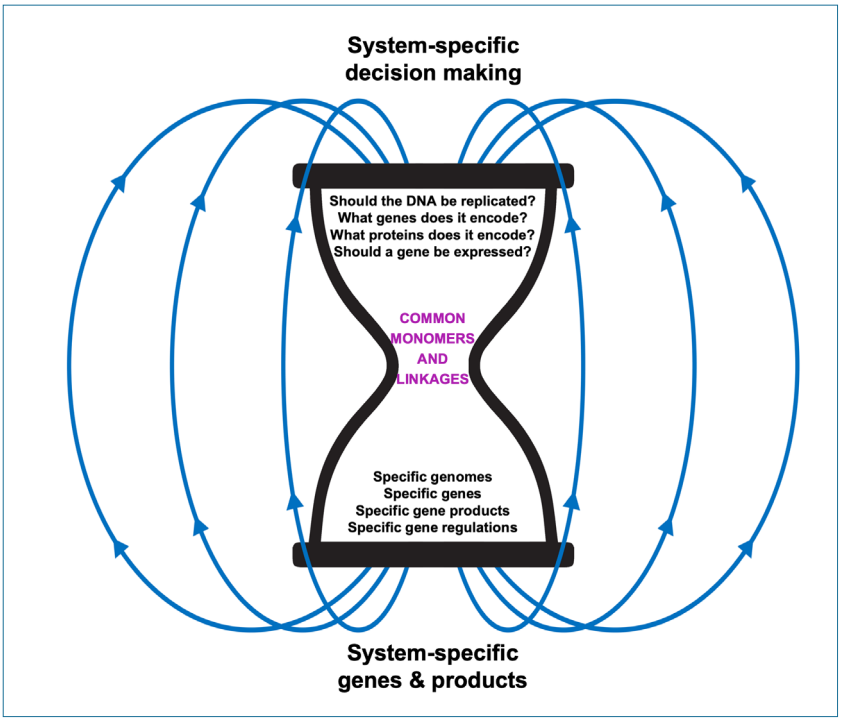


**Figure 2.** An Information transfer hourglass**.** Top: system-specific decision making; bottom: system-specific genes and gene products; middle: non–system-specific monomers, their linkages, and linking chemical reactions. (After figure 8 in ref. 8.)

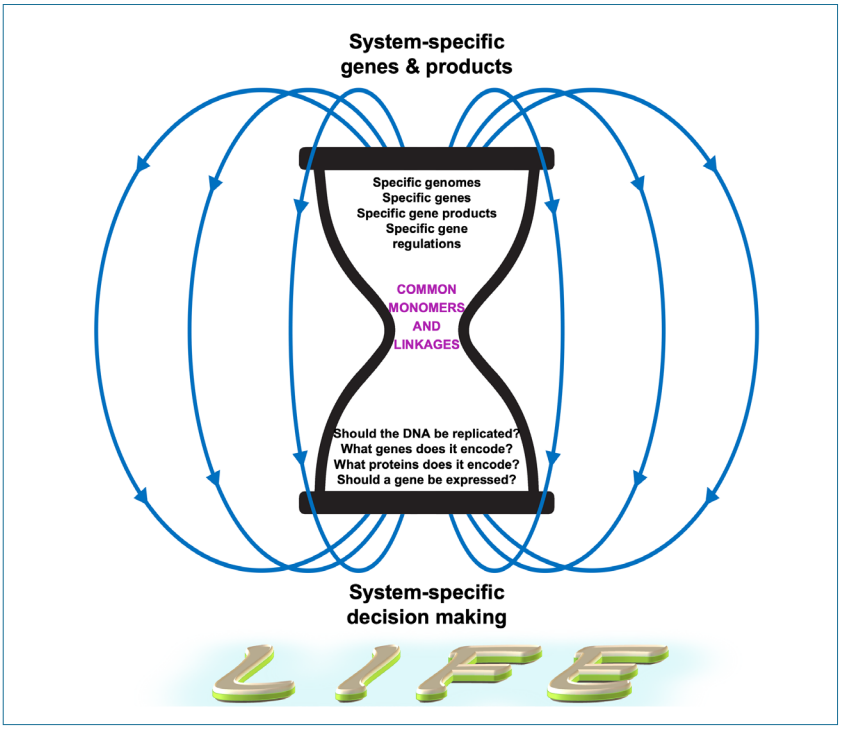


**Figure 3.** An information fountain. Bottom: system-specific decision making; top: system-specific genes and gene products; middle: non-system-specific monomers, their linkages, and linking chemical reactions. Sitting the information fountain on top of 'LIFE' implies that the cell must be initiated with all the decoding equipment in a functioning state.

existence and propagation of cellular life as we know it.

In short, the information flow is circular and directional. Sequence information flows in the direction of DNA to RNA to proteins during gene transcription and translation. Some of the sequence information is lost because not all DNA is gene-coding, and not all RNA, or regions of RNA, code for amino acids within a protein. Meanwhile, episequence information (or cellular and environmental context information) is incorporated during gene transcription and translation, as well as during RNA and protein processing, such as cell-specific alternative intron splicing and protein processing. Therefore, one cannot deduce the genome sequence based solely on the sequences of the RNA or proteins that are encoded by the genome for two reasons. First, there is no reverse gene translation (i.e., there is no molecular machine that can run translation backward). Second, not all RNAs are protein-coding, not all regions of a protein-coding RNA are translated, RNAs can be spliced and edited, and proteins can be spliced and modified—and all in a system-, cell-type-, cell-status-, and environment-specific manner.

## An information fountain

The analogy of the hourglass, even with a lengthy explanation, as above, may give the wrong impression that when the hourglass is flipped, the information transfer would be reversed so that things would go backwards, like the sand in a regular hourglass, so that an mRNA can be generated using its encoded protein as a template.

We think a better representation of the information flow is an information fountain (figure 3). Although the information fountain looks like a flip of the information-transfer hourglass, the concept of the information fountain immediately invokes the image of the directionality of information flow. When all necessary conditions for a water fountain to work are met, water springs upward and then falls, feeding the fountain either as water or as water vapour in the great water cycle. Likewise, when cells are equipped with matching coding and decoding systems and provided all necessary episequence information, the sequence information of the target genes will be transcribed and/or translated. The resulting gene products (either RNA or proteins) then become part of the episequence information of the cell that makes, modifies, or degrades other molecules inside the cell and enables the cell to grow, to reproduce, to differentiate, to migrate, or to die.

Furthermore, just as a water fountain is not a perpetual machine but needs the supply of source water and energy and an intact structural construction, the biological information fountain needs the supply of materials, energy, and an intact cell structure. More importantly, the cell must be alive. Once a cell dies, all information transfer ceases.

## Conclusion

The central dogma of molecular biology formulated by Crick and modified and popularized by Watson has had profound impact not only on biological research but also on our view of life. However, the central dogma's simple concept of information and information transfer fails to account for the molecular reality inside cells. We suggest updating the information to include both sequence and episequence information and updating the central dogma to regulated, dynamic, system-dependent information coding and decoding, which can be symbolized as an information hourglass and, better yet, an information fountain.

## Acknowledgment

I would like to thank the anonymous reviewer for his constructive feedback.

## References

1. Crick, F., Central dogma of molecular biology, *Nature* **227**:561–563, 1970.
2. Crick, F.H., On protein synthesis, *Symp. Soc. Exp. Biol.* **12**: 138–163, 1958.
3. Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M., and Losick, R., *Molecular Biology of the Gene*, 6th edn, Cold Spring Harbor Laboratory Press, 2008.
4. Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M., and Losick, R., *Molecular Biology of the Gene*, 7th edn, Pearson, 2013.
5. Koonin, E.V., Does the central dogma still stand? *Biology Direct* **7**, 27, 2012; p. 1.
6. Tropp, B.E., *Molecular Biology: Genes to proteins*, 4th edn, Jones and Bartlett Learning, LLC, p. 22, 2012.
7. Crick, ref. 1, p. 561.
8. Tan, C.L. and Anderson, E.H., Revising the central dogma: regulated, dynamic, and system-dependent information coding and decoding, *BioComplexity* **2024**:1–21, 2024.
9. Koonin, E.V., Why the central dogma: on the nature of the great biological exclusion principle, *Biol. Direct* **10**, 52, 2015.
10. Camacho, M.P., Beyond descriptive accuracy: the central dogma of molecular biology in scientific practice, *Studies in History and Philosophy of Science Part A* **86**:20–26, 2021.
11. Yockey, H.P., *Information Theory and Molecular Biology*, Cambridge University Press, 1992.
12. Baralle, F.E. and Giudice, J., Alternative splicing as a regulator of development and tissue identity, *Nat. Rev. Mol. Cell. Biol.* **18**:437–451, 2017.
13. Betney, R., de Silva, E., Krishnan, J., and Stansfield, I., Autoregulatory systems controlling translation factor expression: thermostat-like control of translational accuracy, *RNA* **16**:655–663, 2010.