



## The systemic concept of the gene, at age fifteen, and comments on C.N. El-Hani's article 'Between the cross and the sword: The crisis of the gene concept'

### To the Editor:

The concept of the gene is still under hot debate, after about one century of the coining of the term. It was hoped that the advances in molecular biology would be able to produce a clear definition but there was only partial success. This is a message of caution and humility to scientists. A recent report in *Nature* quotes 'a loose definition that could accommodate everyone's demands' - *a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions* - produced by a group of the Sequence Ontology consortium, coordinated by K. Eilbeck (Pearson, 2006). In this definition, I would like to highlight the words 'corresponding to', which points to the relational or functional aspect of the gene, the constant component of all definitions. The gene is a portion (one or more segments) of the genome (including the RNA genomes of some viruses) that corresponds to a product molecule, and both the product and the gene are defined by the system. The 'unit of inheritance' aspect of the definition in Pearson (2006) is the main subject of analysis of the 'crisis' by El-Hani's work (El-Hani, 2007) and needs to be modified also to accommodate the instances of prionic transfers, epigenetic inheritance, maternal effects etc., including the inheritable RNAs with full and long-lasting effects on organisms.

El-Hani's paper gives the reader a full and updated account of the problem and duly presses the point that a definition should serve practical purposes, not attempting the 'unreachable goal' of being 100% predictive of the correct genomic segments, in the examinations guided by present day knowledge, which is not complete. This is the goal of the bioinformatics top-down approach. It is also given credit to the more realistic bottom-up experimental procedures, going in the reverse direction, from the product to the producer. This is the proposal of the systemic concept, acknowledging that the expression of the genomic content is contextual, referring to the cellular processes that respond to internal and external circumstances to search inside its genetic memories for the relevant segments to compose the products at each place and time (see the discussion in El-Hani's paper).

It may be said that scientists will always remain investigating genomes through a double approach: (a) in the top-down mode, making predictions that are only suggestive of putative gene segments, interspersed among non-

coding segments, and then indicating the relevant tests that may or may not be confirmatory; (b) in the bottom-up mode, having a product at hand and then looking for the correspondent genomic segments involved in its synthesis. The latter approach is highly rewarding for learning how the cellular network worked to produce the molecule and builds more knowledge to be utilized for improving the predictive procedures. Both are 'fishing' for the genes and are mutually stimulatory and interdependent.

The definition of the products (RNA or protein) considered relevant to the question of 'what we wish to find the gene for' might be the most difficult part of the task in the bottom-up approach. When we have a defined product it becomes easier to find the gene for it. When the product is a protein, the best option is to restrict it the closest possible to the primary product of translation. If post-translational modifications are included, discussions on where and when to stop with their addition may become endless. In this case, of genes for proteins, all non-translated parts of the genome are excluded from the definition of the gene.

The problem acquires an entirely different shape when the product is an RNA molecule that will not be translated. How 'mature' and at which level of processing should the molecule be, to be considered a functional product for which we should demarcate and specify the gene? There is still a long way to go in this direction, with contributions from both the predictive and the experimental modes of learning.

Some reminders on the expected difficulties arise from knowledge obtained with the genes for proteins. The standard procedures of deriving the genomic sequences corresponding to a polypeptide through the application of the genetic code, or the reverse, of finding long genes which contain unusual segments inside them, that do not fit the standard genetic code, lead to the identification of a variety of non-standard ways of encoding or decoding the genetic information. These cases are strong evidence for the applicability of a systemic concept. The discovery of introns and all further consequences of this, such as genetic sequences inside introns and the varied alternative splicing modes, among others, are already common knowledge, as well as the variant genetic codes (in the few organisms or in the organelles where all proteins are decoded in a way different from the standard), but attention should be given to the various instances of RNA editing and of translational recoding.

These may be considered punctual solutions for coding problems, where mutations altered the genes but, instead of waiting for new mutational compensations or corrective revertants, a systemic solution was adopted. At editing, the genomic sequence (and its immediate transcript) was fixed in a way that is not adequate for decoding through the standard genetic code, but it was made possible to remain through the recruitment of enzymes that modify the RNA, correcting it to make translation possible. At recoding, the mRNA contains sites that are not translatable in the standard way but the correction that makes possible the production of the adequate protein derives from the recruitment or development of unusual decoding mechanisms, either in specific tRNAs, synthetases or ribosomes. The best known cases are those of decoding internal stop codons via systems including suppressor tRNAs, often called the 21<sup>st</sup> and the 22<sup>nd</sup> amino acids (selenocysteine and pyrrolysine, respectively), but the recoding instances already sum to the hundreds (Baranov *et al.*, 2003). It is possible to indicate that some of the recoding could be generally understood as punctual expansions of the genetic

decoding system, in the same way as some of the variant genetic codes could be indicative of reductions.

Systemic plasticity should be recognized and valued. It is strikingly evident in the cases of addition of new genes derived, *e.g.*, from duplications that incorporated mutational variants, or from horizontal transfers. The new genetic piece will only acquire usefulness and meaning to the system when it is adequately embedded in it, through all the necessary regulatory, stabilizing and decoding connections. This process requires that the preexisting hookups of the host system's genes are plastic enough, never frozen, optimized or specific to very high levels, so that they can be shared by the new genetic segments. At later stages, adequate specificities may be developed for the new genes.

*Romeu Cardoso Guimarães*

## References

- Baranov PV, Gurvich OL, Hammer AW, Gesteland RF and Atkins JF (2003) Recode 2003. *Nucleic Acids Res* 31:87-89.
- El-Hani CN (2007) Between the cross and the sword: the crisis of the gene concept. *Genet Mol Biol* 30:297-307.
- Pearson H (2006) What is a gene? *Nature* 441:399-401.