The role of information in cell regulation

Keith Baverstock
Department of Environmental science
University of Eastern Finland
Kuopio Campus
Finland

Abstract

The organised state of living cells must derive from information internal to the system; however, there are strong reasons, based on sound evidence, to reject the base sequence information encoded in the genomic DNA as being directly relevant to the regulation of cellular phenotype. Rather, it is argued here, that highly specific *relational* information, encoded on the gene products, mainly proteins, is responsible for phenotype. This regulatory information emerges as the peptide folds into a tertiary structure in much the same way as enzymic activity emerges under the same circumstances. The DNA coding sequence serves as a data base in which a second category of relational information is stored to enable accurate reproduction of the cellular peptides. In the context of the cell, therefore, information is physical in character and contributes, through its ability to dissipate free energy, to the maximisation of the entropy of the cell according to the 2nd law of thermodynamics.

Key words: information; cell regulation; entropy; free energy; 2nd law of thermodynamics; evolution**.**

Introduction

Living cells are pre-eminently an organised state of matter and where the origin of that organisation lies is a fundamental question for biology. Cell and molecular biology is almost exclusively based upon the assumption that the organised state is derived from information contained in the genotype, which in turn is contained in the nucleus of the cell. There is no *a priori* reason for this assumption; it is largely a product of history. As it is generally accepted that biological properties, function and morphology, i.e., phenotype, are predominantly derived from the properties of proteins (but also some small RNAs), an equally acceptable assumption might have been that the information responsible for cellular organisation resided in those molecules, placing emphasis on the role of cytoplasm rather than the nucleus. In this note I propose that the DNA base sequence information cannot be the information that determines cell phenotype and therefore an alternative source needs to be identified.

Biological information, to be meaningful, requires a semantic component (Jablonka, 2002); that is, it must be richer in its content than Shannon information. The base sequence of DNA can be represented as Shannon information; however, that part (less than 5%) which codes for gene products, effectively determining the concatenation of amino acids in a specific order to produce a defined peptide, can clearly be regarded as a different and richer category. In this case a triplet of DNA bases definitively *relates* to a specific amino acid. Sequence information in DNA, therefore, serves as a template for producing peptides, the precursors of the active gene products, proteins, and is a component of the information inherited on cell division. An important question that arises is the extent to which this relational sequence information is capable of defining biological function. The Central Dogma (CD) says that it is, but the CD was seriously challenged in 2001 when the Human Genome Project showed that only some 25,000 coding sequences were responsible for in excess of 100,000 proteins in the human cell (Carninci, 2008) – that is, each coding sequence is capable of giving rise to on average at least four proteins.

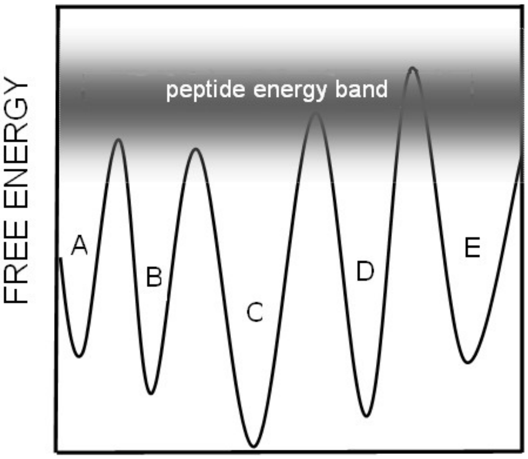

Figure 1: Free energy dependence of tertiary structures A to E derived from the same peptide

There is strong evidence that proteins *relate* to one another as well as to specific coding sequences on DNA, when, for example, they initiate transcription. Organelles, such as ribosomes and cetrosomes, are self-organised and ribosomes can even be functional *in vitro* (Traub and Nomura, 1969). Given the critical importance of the correct function of the ribosome it must be assumed that the interactions between the components (proteins and RNA) are highly specific and, thus, critically dependent on relational information. Additionally, the existence of empirical protein interaction networks (PINs) (Kohn, 1999) indicates that the cellular phenotype relies on proteins working together in the processes that gives rise to cellular functionality. It is reasonable, therefore, to assume that this organisation relies on relational information. For clarity, sequence relational information is termed type I information and protein-protein relational information is termed type II information. The question arises as to whether types I and II are, in fact, the same information.

Characterisation of cellular information

The CD stipulates that the folding of a peptide to form a protein is determined by the amino acid sequence of the peptide

(Anfinsen's dogma) and thus, if correct, the information responsible for phenotype could be the DNA coding sequence information. However, I maintain that this is not the case for several reasons in addition to the point made above concerning the lack if determinism of the transcribed products with respect to the coding sequence:

a) For large peptides several folding options exist leading to several tertiary protein structures with closely similar energy states and separated by relatively large energy barriers (see figure 1), due to the energy required to "reverse", particularly, the initial folds. Under equilibrium conditions structure C would be the minimum energy state but in non-equilibrium conditions, such as applies in open systems, A, B, D and E are all accessible and stabilised by relatively high energy barriers.

b) The established presence of chaperone proteins that among other functions[1] assist actively the folding of peptides to proteins demonstrates that the peptide sequence does not necessarily contain the information required to *determine* its tertiary structure.

c) In addition, there is evidence that in the eukaryotic cell many proteins are partially denatured (Romero et al., 2004) and only adopt their full tertiary structure on approach to a binding site (Sugase et al., 2007) suggesting that more than one folding option is utilised.

d) The protein folding problem has remained unsolved for several decades. According the group of Annila (Sharma et al., 2009) this failure can be understood if protein folding is regarded as a natural, i.e., evolutionary, process to minimise as efficiently as possible free energy according to the 2nd law of thermodynamics. On this interpretation the protein folding problem is very hard because the evolution entails a non-Euclidian energy landscape and is thus non-deterministic, i.e. violates Anfinsen's dogma.

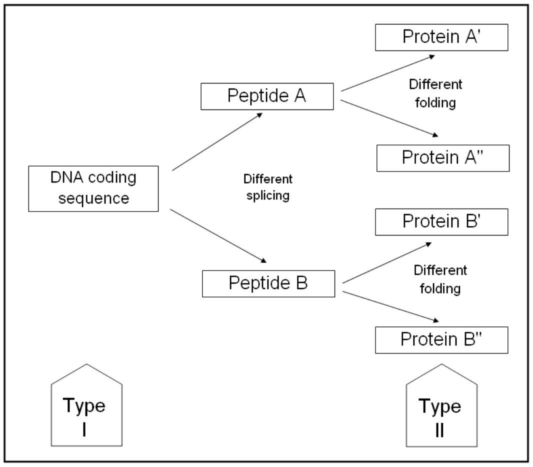

Figure 2: Illustration of how a single coding sequence can lead to several proteins through different splicings and foldings (not necessarily confined to 2).

Taking the above considerations and the degeneracy of gene products per coding sequence together, it must be concluded that the information which is embodied in the right hand side of figure 2 (type II) *cannot* be definitively derived from sequence information on the left hand side of figure 2 (type I). Neither can knowing what binding sites a gene product will recognise provide information on the peptide sequence and thus, the DNA code. The two kinds of information are entirely independent. Neither source of information is useful on its own: phenotype cannot be reliably replicated without the type I information but then the phenotype can only be expressed with the aid of the type II information. Thus, as far as the information (type II) that gives rise to phenotype is concerned the genotype is *empty*.

[1] There is much controversy regarding the various roles of chaperones. One role seems to be to relieve congestion in the confined environment of the cell to stop peptides aggregating and another to provide a space within which a single peptide is free to fold. However, the complexity of the roles of chaperones does not preclude active assistance with folding.

Discussion

I have argued at this workshop and previously, that it is type II information, in the form of "relations" or "rules of engagement" (Baverstock and Rönkkö, 2008) between gene products that is responsible for the cellular phenotype. Information, in this case, is physical in nature – a part of the molecular properties of the gene products (although the exact nature of it has still to be identified) and thus its semantic content is a given. An analogy is, for example, a written notice in the Finnish language offering free coffee which only has meaning for certain recipients who understand the Finnish language; for others there is still information (in the Shannon sense) but it is meaningless: the Finnish language speaker can act upon the information to obtain a free cup of coffee whereas the non-speaker cannot. Thus, all the essential components of a communication, namely, sender, recipient, meaning and interpretation are present in the material application of the relation. Indeed, according to Karnani et al (Karnani et al., 2009) information can be viewed as a free energy transduction process which increases the entropy of the receiver. Thus, the relations that give rise to phenotype are contributing to the maximisation of the entropy of the system.

Increased entropy, as commonly understood in the context of Boltzmann's treatment of thermodynamically *closed* systems, is associated with increased disorder. However, this need not be the case for *open* systems (Sharma and Annila, 2007). The 2nd Law stipulates that in all free energy driven natural processes, entropy increases and this is in effect saying that the disparities in free energy (for example, between a cell and its environment) will be minimised in the least time possible (Annila, 2010). If, as is clearly the case, that minimisation can occur more efficiently through a state of organised energy transduction than through a disordered state, the organised state will be selected. Thus, on this basis information plays a critical physical role in biology.

In addition to the role of type II information in determining phenotype there are two other contexts where it is important; between gene products (transcription factors) and DNA and for the self-organised cellular components, ribosomes, centrioles etc.. Thus, the source of information that has so far been considered of secondary importance, if it has not been neglected altogether, would appear to be of equal importance to sequence information in modern cells regardless of whether regulation is seen in terms of the independent attractor model (Baverstock and Rönkkö, 2008) or the genetic regulatory network model (Huang, 2009).

At the workshop it was an almost universal assumption that causality in cellular regulatory processes runs from genotype to phenotype. For example, if I interpret him correctly, Stig Omholt's causally-cohesive genotype-phenotype (cGP) model assumes this. However, the cGP approach could not readily account, in terms of allele frequencies, for the strong offspring–parent resemblances observed (Gjuvsland et al., 2011). To achieve the high levels of additive variance implied by such resemblances it was necessary to apply constraints to the model parameters.

The argument outlined above provides another interpretation of why the observed resemblances between parent and offspring are not more readily accounted for. In terms of the type II information, governing gene product interactions, the genotype does not contain the relevant information – it cannot therefore predict phenotypic properties. Support for this contention comes in a recently reported monozygotic twin study (Roberts et al., 2012). This showed that genomic sequence was a poor predictor of predisposition to 19 out of the 24 common diseases examined.

In respect of information the genotype only serves the purpose of storing the base sequence data that enables the gene products to be synthesised with high integrity. This role is of course vital: from the thermodynamic perspective a modern cell is the culmination of an evolutionary process, measured in billions of years, that has refined, through natural selection, its ability to increase its entropy, or reduce the free energy disparity at its interface with its environment. Only by ensuring accurate replication of their components can cells maintain and build upon, i.e., further evolve from, their current state.

Acknowledgements

I am grateful to Arto Annila and Tom Melham for productive discussion.

References

Annila, A. (2010) The 2nd Law of Thermodynamics Delineates Dispersal of Energy. *Int. Rev.Phys.* **4,** 29 - 34.

Baverstock, K. and Rönkkö, M. (2008) Epigenetic regulation of the mammalian cell. *PloS One* **3,** e2290.

Carninci, P. (2008) Hunting hidden transcripts. *Nat Methods* **5,** 587-9.

Gjuvsland, A. B., Vik, J. O., Woolliams, J. A. and Omholt, S. W. (2011) Order-preserving principles underlying genotype-phenotype maps ensure high additive proportions of genetic variance. *J Evol Biol* **24,** 2269-79.

Huang, S. (2009) Reprogramming cell fates: reconciling rarity with robustness. *BioEssays* **31,** 546-560.

Jablonka, E. (2002) Information: Its Interpretation, Its Inheritance, and Its Sharing. *Philosophy of Science* **69,** 578 – 605.

Karnani, M., Pääakkänen, K. and Annila, A. (2009) The physical character of information. *Proc. R. Soc. A* **doi:10.1098/rspa.2009.0063**.

Kohn, K. W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell* **10,** 2703-34.

Roberts, N. J., Vogelstein, J. T., Parmigiani, G., Kinzler, K. W., Vogelstein, B. and Velculescu, V. E. (2012) The predictive capacity of personal genome sequencing. *Sci Transl Med* **4,** 133ra58.

Romero, P., Obradovic, Z. and Dunker, A. K. (2004) Natively disordered proteins: functions and predictions. *Appl Bioinformatics* **3,** 105-13.

Sharma, V. and Annila, A. (2007) Natural process--natural selection. *Biophys Chem* **127,** 123-8.

Sharma, V., Kaile, V. R. I. and Annila, A. (2009) Protein folding as an evolutionary process. *Physica A* **388,** 851 - 862.

Sugase, K., Dyson, H. J. and Wright, P. E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447,** 1021-1025.

Traub, P. and Nomura, M. (1969) Structure and function of Escherichia coli ribosomes. VI. Mechanism of assembly of 30 s ribosomes studied in vitro. *J Mol Biol* **40,** 391-413.