Correspondence

Rosetta Stone proteins: "chance and necessity"?

Reiner A Veitia

A response to **Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions** by AJ Enright, CA Ouzounis. *Genome Biology* 2001, **2:**research0034.1-0034.7

Address: UFR de Biologie et Sciences de la Nature, Université Denis Diderot/Paris VII, Immunogénétique Humaine, Institut Pasteur, 75724 Paris, France. E-mail: rveitia@pasteur.fr

Published: 8 January 2002

Genome ***Biology*** 2002, **3(2)**:interactions1001.1–1001.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/2/interactions/1001

The field of predicting protein-protein interactions has been active for over two decades, but *in silico* methods have become prominent in the last three years with the expansion of analyses at a genomic scale [1,2]. The rationale of one of the computational methods is as simple as it is elegant. Two polypeptides A and B in one organism are likely to interact if their homologs are expressed as a single polypeptide AB in another. The latter polypeptide (AB) is called a Rosetta Stone protein, as it contains information about both A and B. Marcotte *et al.* [1] have proposed that fusion to form a single polypeptide reduces the entropy of dissociation of A and B. The result is a huge increase in the local concentration of A with respect to B. A recent paper in *Genome Biology* describes the effort of Enright and Ouzounis [3], who carried out a vast analysis of genes of the Rosetta Stone type over 24 genomes, including eukaryotic genomes. They uncovered many new 'composite' or 'Rosetta' proteins, many of them contributed by eukaryotes. Here, I provide simple arguments to suggest that including eukaryotic sequences in the analyses may increase the robustness of predictions made using the Rosetta Stone approach.

In prokaryotes, transcription and translation are coupled, and functionally related genes are clustered. Dandekar *et al.* [4] compared nine prokaryotic genomes and noticed a poor conservation of architecture of operons (clusters of co-transcribed functionally related genes) - but what is an operon in one organism may, in another, be a regulon (several co-regulated operons or suboperons). They noticed that a number of gene pairs were highly conserved, taking this as evidence for direct physical interactions between the corresponding gene products, rather than a reflection of co-regulation resulting from functional coupling (see also [5]). Given that a biochemical function in many cases depends on the action of a multimeric complex, a correlation between co-regulated and interacting proteins is to be expected (corresponding to a proportion of the positive hits in the approach of Dandekar *et al.*).

The rationale of the Rosetta-type search seems to be far more robust, but the existence of a Rosetta protein is not always proof of protein-protein interactions. The operon can be considered as a selfish cassette of DNA that can confer a selective advantage under certain conditions. The operon is therefore a gene cluster that has been assembled by deleting 'uninteresting' intervening sequences and can be spread by horizontal gene transfer to many recipient genomes [6]. From a minimalist perspective, it is conceivable that deletion of an intervening sequence between two adjacent genes may lead to an in-frame fusion of the open reading frames (ORFs). If folding of the fused proteins is not altered, this is a way to co-regulate gene expression as efficiently as an operon does for separate ORFs. Thus, fusion events may reflect an alternative strategy of co-regulation and not direct physical interactions. This might explain, at least in part, one surprising result of Enright and Ouzounis [3]: when they tried to validate their predictions using the results of a yeast two-hybrid experiment on a genomic scale, only one case found validation. This may also reflect, as the authors notice, the extremely high number of false-positive hits of the two-hybrid method. Consider also that *Mycoplasma genitalium* (with a 580 kb genome containing 479 ORFs), which has a genome smaller than that of *Mycoplasma pneumoniae* (816 kb, 677 ORFs), nevertheless contains 15 Rosetta proteins whose *M. pneumoniae* homologs are encoded by split genes. The reverse comparison shows that *M. pneumoniae* has only four Rosetta proteins when the reference genome is that of *M. genitalium*. Although this does not preclude the possibility of physical interactions between the putative partners, it can be used as a circumstantial argument to

suggest that reductive evolution may push towards gene fusion for the sake of economy.

Although in prokaryotes fusion events may in some instances reflect co-regulatory strategies, the introduction of eukaryotic genomes into searches for Rosetta proteins may help improve the robustness of predictions, especially when the hits involve organisms from different kingdoms of life (eukaryotes versus bacteria and archaea). Instead of using 'energetic' arguments (changes in entropy, ΔS, or in the Gibbs' free energy, ΔG), we can use a simple mass-action approach to justify this. Consider, for example, a dimerization reaction characterized by the equilibrium constant K, such that

$$A + B = AB, \qquad K = \frac{AB_{eq}}{A_{eq}.B_{eq}}$$

where A_{eq}, B_{eq} and AB_{eq} are the concentrations at equilibrium. If A_o and B_o are the initial concentrations of A and B, the constant can be expressed as

$$K = \frac{AB_{eq}}{(A_o - AB_{eq})(B_o - AB_{eq})}$$

This formula is an ordinary quadratic equation, and we can find the value of AB_{eq} analytically. Imagine now that we have a cell that swells (expanding from the volume of a prokaryotic cell to that of a eukaryote) without changing the input concentration amounts of each subunit. A modest increase of the volume, say 1,000 times, translates into a dramatic decrease in the amounts of AB. For instance, for K = 10^8 M^{-1} and $A_o = B_o$ = 1 μM, a 1,000-fold dilution leads to over 10,000 times less AB at equilibrium (see Figure 1a). Even worse, if the starting concentration of A and B is 1 nM, the same dilution leads to over 800,000 times less AB at the end. Increasing K may help to overcome this problem. If now K = 10^{12} M^{-1} for $A_o = B_o$ = 1 μM, we will have only 1,000 times less AB (which is a proportional change with respect to the dilution factor). However, if $A_o = B_o$ = 1 nM, 1,000 times less of AB is obtained for K > 10^{15} M^{-1}.

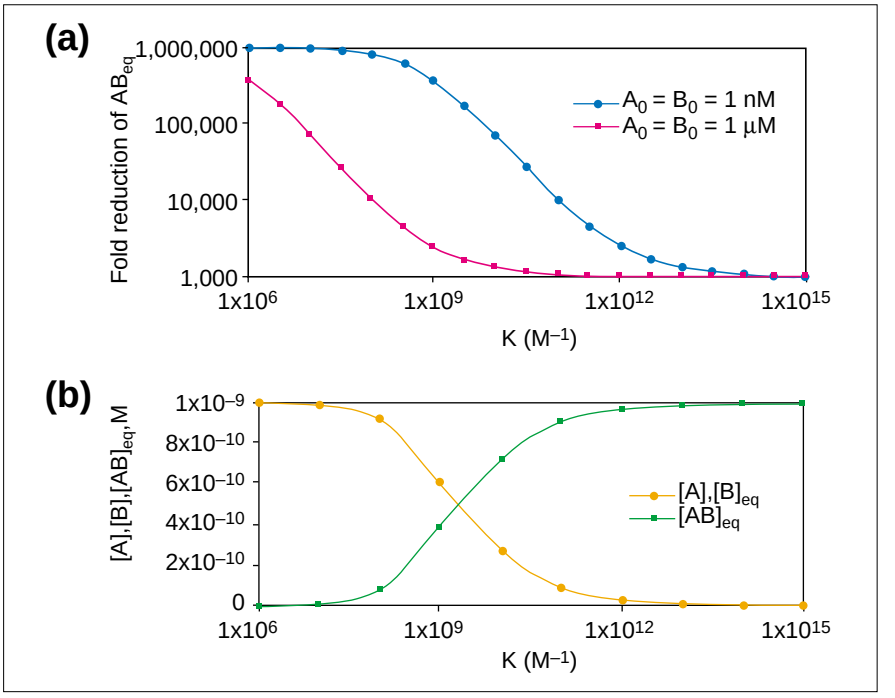

Figure 1
(a) The relationship between the fold reduction of the equilibrium concentrations of AB (in the dimerization reaction discussed in the text) and the association constant K following a 1,000-fold dilution of A_0 and B_0. Two different initial concentrations of A and B have been considered (1 μM and 1 nM). Calculations were made using the analytical solution of the equation: fold reduction = (AB_{eq}(for A_0,B_0))/(AB_{eq}(for A_0/1,000,B_0/1,000)). Note that both axes are logarithmic. It is easy to see that the differences in the yield of AB after dilution depend on the initial input concentrations and are enormous for low K values (weak complexes). These differences, however, decrease as K increases (tight complexes). Notice that for lower initial concentrations of A and B the curve moves towards the right, and higher K values will be required to obtain a similar fold-reduction relative to the situation where the initial monomer concentrations are higher. Fusion of A and B is intuitively equivalent to increasing K to infinity or to enormously increasing the amounts of A able to 'see' B. There is a link between K and classical thermodynamic parameters (such as those evoked in [1]) but these parameters are less immediate than the mass-action arguments and are of minor importance for understanding here. **(b)** The equilibrium concentration of A (or B) and AB as a function of K (M^{-1}) for $A_0 = B_0$ = 1 nM. It is clear that for low K the association reaction is very inefficient. As K increases, the concentration of free monomers (physicochemically required to fulfill the constraints of the mass action law but biologically useless) decreases. Fusion of A and B is the best option available to avoid monomer wasting and the diffusional problems mentioned in the text.

It is clear that increasing the affinity (K) between the partners enhances dimer formation even at low input doses of monomers (Figure 1b). The increase in K required can be enormous, however, and even in the case of an irreversible (ultra-tight) dimerization, it will always be limited by the translational diffusion of the partners. If the cell requires a certain amount of a product at a certain moment, this may be unattainable, from a kinetic point of view, with low monomer concentrations. This is especially critical in eukaryotes, where co-regulated genes are not physically linked and where transcription and translation take place in different compartments. If similar levels of AB activity were needed by both types of cells (that is, a small or prokaryotic cell and a large or eukaryotic cell), three main non-exclusive alternatives are possible: first, a proportional increase in the input molar concentrations in the large cell; second, the introduction of compartments in

the large cell; and third, fusion of the interacting partners. The first strategy is not parsimonious because, in our hypothetical example, a 1,000-fold increase in the initial concentrations of all interacting partners must be guaranteed (involving a 1,000-fold increase of 'biologically useless' free monomers, that is, A_0-AB_{eq}, B_0-AB_{eq}). This might also imply drastically slowing down the turnover of the proteins so as to allow their accumulation. The second strategy is more advantageous. Note, however, that in most cases independent polypeptides must diffuse across the cytosol (site of synthesis) to reach their compartments. Since the compartments and/or organelles are big targets and able to sequester the proteins, they may relieve the diffusion problem, whose consequence is only kinetic. If the cell needs high concentrations of the AB complex in a short period of time, however, increasing the input concentration of monomers will be required (in spite of the 'relief' provided by the organelles); if time is not a problem, by contrast, the cell can 'wait' until the organellar concentrations of monomers and/or complexes are the correct ones. The strategy of gene/protein fusion is the most parsimonious and kinetically advantageous. It dramatically helps to diminish the amount of transcribed and translated products required to attain the desired levels of functional activity. This is clearly beneficial in terms of time and energy consumption, and can also be applied to proteins sorted to specialized compartments (no diffusion of monomers is required to enable them to meet inside the compartments or within membranes).

In the case of enzymes, we can also evaluate the advantages of gene or protein fusion from the perspective of the chemical reactions they catalyze. In prokaryotes, after translation, whether the gene products interact physically or not, they are all produced in relative proximity. In eukaryotes, protein fusion to yield polyproteins able to catalyze successive steps of a metabolic pathway provides a great advantage compared to producing independent polypeptides. Note that the partition of the cellular volume into organelles also enhances metabolic efficiency and that, perhaps, the existence of organelles is linked to improving metabolic processes rather than to relieving the protein diffusion problem evoked above. This is reminiscent of the notion of metabolic channeling, used to describe the restricted flow of substrates and products in multienzyme systems (substrates and/or products are passed from one active center to another). It has been argued that free diffusion is sufficiently rapid to obviate the need for channeling [7]. Again, this is easily applicable to prokaryotes. In eukaryotes, however, large cytosolic volumes may result in a greater need for channeling.

It is safe to assume that most Rosetta proteins conserved across eukaryotes and prokaryotes are responsible for the 'core' metabolism (intermediary and basic information transfer, as defined in [8]). Consistent with this, the comparison of *Drosophila* with other organisms [9] shows that, in almost all cases where functional annotation is known, the Rosetta components are involved in the core metabolism. Exploitation of these results might aid understanding of how simple organisms work. Even this would be a big achievement, which would in turn help us to understand more complex eukaryotic systems. From the perspective outlined here, eukaryotic Rosetta proteins are likely better to reflect protein-protein interactions (producing fewer false positives) than those found outside Eukarya (many of which are also relevant). On the other hand, Rosetta proteins specific to eukaryotes might reflect the modular nature and the combinatorial design of many eukaryotic components (complex transcription factors, signal transduction molecules and molecular adaptors). It is conceivable that nature could have evolved a huge combinatorial panoply of enzymes with a limited set of interacting generic domains (cofactor-binding and catalytic domains), but in fact selection has favored a 'copy-and-paste' strategy, allowing the multiplication of domains that appear today as fusion products [10]. The results of this copy-and-paste process can be grouped into two main classes: a primordial scenario leading from several domains constituting several peptides to several domains combined into one polypeptide, and a sophisticated scenario leading from several genes encoding several polypeptides to one gene for one polyprotein. The evolutionary path followed from one state to the other is, however, largely unknown.

Acknowledgements

I thank Sandrine Caburet for helpful discussions.

References

1. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285:**751-753.
2. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402:**86-90.
3. Enright AJ, Ouzounis CA: **Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions.** *Genome Biol* 2001, **2:**research0034.1-0034.7.
4. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23:**324-328.
5. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. **The use of gene clusters to infer functional coupling.** *Proc Natl Acad Sci USA* 1999, **96:**2896-2901.
6. Lawrence J: **Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes.** *Curr Opin Genet Dev* 1999, **9:**642-648.
7. Welch GR, Easterby JS: **Metabolic channeling versus free diffusion: transition-time analysis.** *Trends Biochem Sci* 1994, **19:**193-197.
8. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, *et al.*: **Comparison of the complete protein sets of worm and yeast: orthology and divergence.** *Science* 1998, **282:**2022-2028.
9. **All Fuse** [http://maine.ebi.ac.uk:8000/services/allfuse/]
10. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli.*** *J Mol Biol* 2001, **311:**693-708.