Is optimal gene order impossible?

Juan F. Poyatos[1] **and Laurence D. Hurst**[2]

[1]Evolutionary Systems Biology Initiative, Structural and Computational Biology Programme, Spanish National Cancer Centre (CNIO), Melchor Fernández Almagro 3, 28029 Madrid, Spain
[2]Department of Biology and Biochemistry, University of Bath, Somerset BA2 7AY, UK

Recent evidence suggests that yeast genes encoding proteins that are present in the same protein complex tend to be linked and to be co-expressed. More generally, we found that genes that are close to each other in the protein interaction network tend to be linked more often than expected and are often co-expressed. Unexpectedly, we found that linked genes in network proximity have unusually high recombination rates. Because high recombination rates are associated with high rates of genome re-organization, our findings might explain why the clustering of genes in proximity in the network is such a weak effect: there could be a co-evolutionary cycle of physical linkage for co-expression, upwards modification of the recombination rate and concomitant break-up of a cluster. Under such a model an 'optimal' gene order is never stable.

Introduction

There is increasing evidence that gene order in eukaryotic genomes is not random [1]. Why might this be? At least two models have been proposed to explain this phenomenon, but the relative importance of each is unclear. First, genes might cluster to enable co-regulation, mediated either on the small scale (e.g. bidirectional promoters [2]) or on a broader scale (e.g. chromatin-mediated regulation [3,4]). This idea is supported by the finding that genes that are similarly expressed are found in clusters (e.g. genes that are co-expressed [5–7] and/or broadly expressed [8,9] and/or highly expressed [10,11]). Second, genome proximity might evolve to reduce the recombination rate between specific genes [12,13]. Although this idea is well supported by evidence from rare supergene clusters [1], the generality of the model is unclear. Recent evidence that essential genes cluster in regions of low recombination [14] is consistent with this idea [15].

Under both models one may consider the process of the evolution of clustering in one of two different modes: (i) a mode in which a re-arrangement is selectively favoured at the instant of its creation; or (ii) a mode in which the re-arrangement need not be immediately advantageous but subsequent evolution ensures that the re-arrangement is likely to be preferentially preserved over time. In the co-regulation model, chromatin-mediated effects can provide an immediate increase in the level of co-regulation (e.g. Ref. [4]). By contrast, as Lawrence has argued [16], for alternative modes of regulation (operon structures, bidirectional promoters) it might be more parsimonious to suppose that the evolution of co-regulation can only occur after the evolution of genome proximity. Likewise for the recombinational model, if the epistatically interacting variants are sufficiently common (e.g. held by balancing selection) then an immediate advantage is possible [17,18]. However, if the variants are held by mutation selection equilibrium then an immediate advantage is either unlikely or weak. However, after the initial evolution of genome proximity the alleles will approach equilibrium, at which point selection for reduced recombination is possible. This alone could lead to an excess of such gene pairs because low recombination rates are associated with low rates of re-arrangement [19].

Recent work [20] has shown that in the yeast *Saccharomyces cerevisae*, genes in the same protein–protein complex tend both to cluster in the genome and to be co-expressed, consistent with the first model. Here we ask whether such clusters have low recombination rates as predicted by the recombination–modification model. However, rather than specifically analysing genes encoding proteins involved in protein–protein complexes, we ask more generally about proteins that are close to each other in the protein interaction network according to several graph-based measures.

Network proximity and genome proximity

There is, as expected given previous results [20], a relationship between proximity in the protein interaction network and proximity in the genome. To address this issue, we first selected all pairs of genes that are both neighbours in the genome and present in the protein interaction data set and calculated the shortest distance between the proteins in the network, d. We then computed the mean distance in the network for all such pairs. Significance was assessed by randomization in which the genomic positions of the genes were randomized (see supplementary material). We found that the real genomes have a highly significantly lower mean network distance between genomically adjacent pairs than expected by chance ($P=0.0062$, mean distance observed $=3.99$; mean distance random $=4.03\pm0.01$). Although this difference is significant, it is small (discussed in more detail below).

To clarify the relationship between network proximity and genome proximity we performed two further tests. First, we asked what was the number of close proteins in the network (distance $d<3$) whose genes are adjacent in the genome? This number is greater than that expected by

Corresponding authors: Poyatos, J.F. (jpoyatos@cnio.es), Hurst, L.D. (l.d.hurst@bath.ac.uk).

Available online 27 June 2006

chance ($P<0.0001$, close and adjacent observed $=167$; close and adjacent random $=103\pm10$; far and adjacent observed $=3306$; far and adjacent random $=3357\pm15$; see supplementary material online). In addition, we calculated the generalized clustering coefficient, (C_{ij}; see supplementary material online), a measure of the number of interacting proteins that two given proteins have in common. We then asked do genes that are neighbours in the genome have a greater number of interactors in common than expected? Significance was again determined by randomization. We found that neighbours in the genome have a greater than expected number of common interactants ($P<0.0001$, mean clustering observed $=0.019$, mean clustering random $=0.011\pm0.001$). These tests support the view that the network neighbourhood is related to the genome neighbourhood.

These results might have a trivial explanation: we might just be sampling tandem duplicates. To eliminate this possibility, we removed one of the two genes from the tandem pair (i.e. set its connectivity to zero), re-assembled the network and recalculated the relevant parameters. All our results remain significant (Table S1 in supplementary online). Do these results reflect the genome proximity of genes involved in the same protein–protein complex, as previously described [20], or do they reflect a more general phenomenon? To examine this, when analysing a given pair of genomically neighbouring genes, we randomly assigned a connectivity value of zero to one of the two genes if both belonged to the same complex. We then re-assembled the network and repeated the analysis. All results remain significant (Table S2 in the supplementary material online and Table S3 where tandems and complexes were controlled together), suggesting that the effects are generally true for protein–protein interactions and not just a feature of a particular subclass. However, we should be cautious about this result because, owing to annotation errors, stable complex proteins might be still in the data set of putative non-stable complex proteins. Finally, we considered errors in the data by analysing a subset of the interactions whose quality has been assessed by different computational methods (supplementary material online).

Modules and genome proximity

Because the module is considered to be a potentially important level of organization between the gene and the phenotype [21,22], it might also be instructive to ask whether members of the same module are located closer to each other in the genome than expected by chance. The difficulty here is in defining a module, with different modules extracting different biological features. We employed three different definitions (see supplementary material online). We found that in fully connected modules (i.e. cliques), when members of a module appear on the same chromosome, they tend to be more tightly linked than expected by chance (Table 1). Finally, because modules embedded in networks are often associated with protein complexes, we used a list of complexes to determine whether their members are linked. Confirming the previous report [20], we found that members of small stable complexes are linked.

Table 1. Linkage of members of a given module[a,b]

Module type	Significance	Module type	Significance
Clique–3	$P = 0.0158$	Clique–4	NS
Clique–5	$P<0.0001$	Clique–6	$P = 0.0003$
Clique–7	$P<0.0001$	Clique–8	NS
SPC	NS	OVE	NS
Small–COM	$P<0.0001$	Large–COM	NS

[a] Abbreviation: NS, not significant.

[b] For a given module definition, we questioned whether members of the same module are more closely located in the genome than expected by chance. We used three definitions, cliques (clique–*n*, where *n* is the corresponding size), super paramagnetic clustering modules [29,30] (SPC) and overlap modules [22] (OVE) (for more details, see supplementary material online). Only cliques had significant clustering. Note that we considered that all module members should be in the same chromosome in the case of cliques of size three. We also measured the genome proximity of a set of small (small-COM <15 components) and large stable complexes (large–COM, ≥15 components). Small stable complexes tended to be located close together in the genome.

Co-regulation and low recombination models

We can also confirm that the genes in proximity in the network and in proximity in the genome tend to have a greater level of co-expression than those that are equally close in the genome but more distant in the network. To examine the possibility that clustering is advantageous as a means to ensure co-regulation, we calculated the distance between the proteins in the network for each pair of neighbouring genes in the genome. We also calculated the correlation between the expression levels between the genes. We observe that genes that are close to each other in the network ($d<3$) have more examples of positive correlation in co-expression that those more distant in the network ($P=0.0118$, Figure 1a). This is neither a result of tandem duplicates nor is it solely a result of the co-expression of genes found in the same protein–protein complex (Table S5 supplementary material online). This supports the idea that genomic proximity co-evolves with network proximity, mediated by selection on co-expression.

Do genes in proximity in the network and in the genome have low recombination rates? We did the same analysis as described above, this time measuring the recombination rate between the genes. To estimate the recombination rate, we made use of high-resolution data on meiotic double-strand DNA breaks [23]. The recombination rate of a pair of adjacent genes was computed as the mean of the rates for each member of the pair. In addition, we benchmarked the reliability of the data as a surrogate for recombination rate by comparing it with >40 years worth of accumulated recombination measures. The double-strand break data is a reliable measure showing dips at all centromeres and strongly correlates with the observed rates per unit of genetic distance (supplementary material online). Unexpectedly, we observe that genes that are close to each other in the network have greater recombination rates than those more distant in the network ($P=0.0213$, Figure 1b). This is not because of tandem duplicates nor is it restricted to genes found in the same protein–protein complex (Table S5 supplementary material online). These results are contrary to the prediction of the recombinational model, which proposes that physical linkage is a means to ensure close genetic linkage.

Is the high co-expression rate and high recombination rate specific to those genes that are both close in the

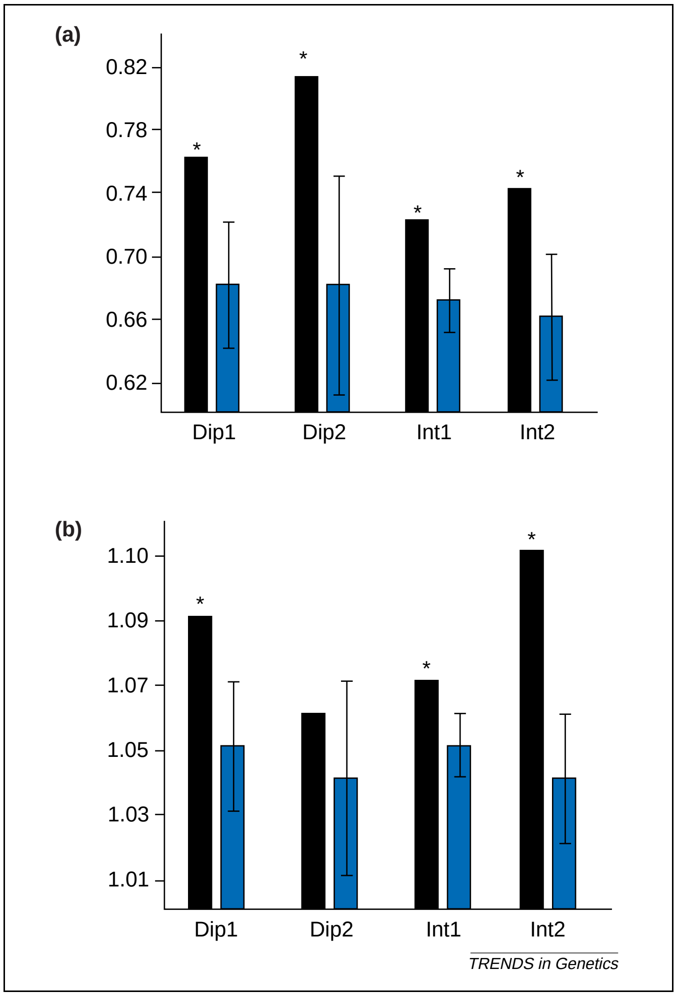

Figure 1. Co-expression and recombination rate of proteins that are close in the network ($d<3$) and adjacent in the genome. **(a)** The ratio of positive co-expression to the number of close and adjacent genes. (**b**) The mean recombination rate. Black bars denote the observed values, whereas blue bars denote the corresponding null average values ± one standard error. Different bars correspond to different data sets (asterisks representing significant values, $P\leq0.05$): Dip1, DIP data set; Dip2, DIP–core; Int1, the so–called TopNet data set; Int2, a subset of the DIP proteins that are close(far) and adjacent in the genome that are also found as close(far) and adjacent using the TopNet (Int1) data set (see supplementary material for further details).

genome and in the network or is it a general feature of those close in the network? Were the former the case, we would expect that those pairs of genes that are close to each other in the network ($d<3$) and adjacent in the genome should experience greater recombination rate and co-expression than the whole set of short distance genes (adjacent or not). Adjacent genes have greater recombination rates ($P=0.0052$) and greater co-expression ($P=0.0004$) than expected by chance were genes drawn from the full set of short distance genes. Could there be other reasons for the high recombination rates found? Genomic regions differ in the recombination rate, perhaps gene pairs in proximity in the network are all found in regions of high recombination. To test this idea, we compared the mean recombination rates of any such pairs with those formed by their immediate neighbours (e.g. we have four genes in a row in the genome A, B, C, D; we compared the recombination rate of genes B and C, which are close to each other in the net, with that of the pair A and D). The mean difference between recombination rates for all such pairs is greater than expected by chance ($P=0.014$, randomly sampling from all adjacent pairs in the genome; i.e. distance independent). Finally, because the recombination rate is greater for highly expressed genes or those with longer intergene distance, it is also necessary to analyse the contribution of these two factors. This does not affect the results because genes that are close to each other in the network and adjacent in the genome are neither more highly expressed ($P=0.88$) nor located further apart ($P=0.76$) than expected by chance.

Concluding remarks

Our results suggest a simple model, in which selection favours the co-expression, to some degree, of proteins that are in proximity in the network. This might be to ensure appropriate balance of members of a complex [24,20] but, as shown here, it is unlikely that this is the full explanation. Movement of genes around the genome, by whatever means, will occasionally bring two genes close together that are in proximity in the network. Some concomitant level of co-expression, perhaps mediated by chromatin level effects (much as inserted genes can assume the expression profile of domains into which they insert [4]) might be associated with this genomic proximity. Alternatively and/or additionally, proximity might permit the evolution of mechanisms that increase the level of co-regulation; for example, by the evolution of bidirectional promoters. Furthermore, if the genes are in the same orientation, they might evolve to enable regulation by the same transcription factor(s) [25]. However, there is no guarantee that the proximity of these genes needs be favoured by selection for recombinational linkage. If, for example, interactions between the genes were on average negatively epistatic, there would be a weak force promoting a local increase in the recombination rate [26,27]. This then could account for the unusually high recombination rates seen for those genes that are close to each other both in the genome and in the network. One can only speculate as to why negative epistasis might be found between interacting proteins, just as one can only speculate as to why positive epistasis might be found. One could consider, for example, that if A and A′ are two versions of a protein that binds to B and B′, that A and B bind well, A′ and B bind adequately, as do B′ and A, but A′ and B′ fail to bind, resulting in negative epistasis.

The high recombination rate that we see could have an additional consequence – the promotion of instability of the ordered genome. In wheat regions of high recombination rate tend to be those that are, over evolutionary time, rearranged [19]. Is the same true in yeast? Taking pairs of neighbouring genes in yeast we can recover their orthologs in *Klyveromyces lactis*. Those that are neighbouring pairs in both species have on average lower recombination rates than those neighbouring only in *S. cerevisiae* (mean recombination rate for all pairs found in *K. lactis*: 1.0380; mean recombination rate for all pairs found as linked in *K. lactis*: 1.0096, $P=0.0029$; randomizing over all conserved pairs, see supplementary material online for more details). Not surprisingly, we found that the gene pairs that specify proteins close in the network ($d<3$) are typically only conserved as a pair if, unusually, they have a low recombination rate. Only 21 of the 88 adjacent and close pairs in *S. cerevisae* with orthologs in *K. lactis* are also found to be adjacent in the *K. lactis* genome. The mean recombination rate for these 21 is 0.95, which is much lower than that expected were they a random selection (mean of 21 pairs randomly selected 1.08 ± 0.05, $P=0.0022$).

Consequently, this result suggests a model in which gene order is forever in flux. Selection favours certain gene pairs to reside in proximity because it permits a greater level of co-expression than is otherwise possible. This, in turn, favours modifiers that increase the level of co-expression but also favours those that increase the level of recombination, if, for example, epistasis between the genes is negative. In the long term, the increased recombination rate causes, we suggest, gene order to be disrupted. A stable gene order, under such a model, is not possible, with the exception of gene pairs with strikingly high levels of co-expression [28] or those that have low recombination rates [14]. Although our results are, in qualitative terms, consistent with this proposed cycle, whether the effects observed are of an adequate magnitude is unresolved. A tendency for recombination to disrupt otherwise 'adaptive' gene orders might also help address the question: why, in any given genome, only a small proportion of genes show evidence for clustering, although non-random gene order is common in eukaryotic genomes?

Acknowledgements

We thank E. Domany and A. Yitzhaky for providing us with the SPC code. This research has been supported by the Spanish MCyT (Ministry of Science and Technology) Ramón y Cajal Program (J.F.P.) and the UK Biotechnology and Biological Sciences Research Council (L.D.H.).

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2006.06.003

References

1 Hurst, L.D. *et al.* (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5, 299–310

2 Trinklein, N.D. *et al.* (2004) An abundance of bidirectional promoters in the human genome. *Genome Res.* 14, 62–66

3 Robyr, D. *et al.* (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* 109, 437–446

4 Finnegan, E.J. *et al.* (2004) A cluster of *Arabidopsis* genes with a coordinate response to an environmental stimulus. *Curr. Biol.* 14, 911–916

5 Cohen, B.A. *et al.* (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186

6 Williams, E.J. and Bowles, D.J. (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res.* 14, 1060–1067

7 Spellman, P.T. and Rubin, G.M. (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* 1, 5

8 Lercher, M.J. *et al.* (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* 31, 180–183

9 Lercher, M.J. *et al.* (2003) A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* 12, 2411–2415

10 Caron, H. *et al.* (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292

11 Versteeg, R. *et al.* (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 13, 1998–2004

12 Nei, M. (1968) Evolutionary change of linkage intensity. *Nature* 218, 1160–1161

13 Fisher, R. (1930) *The genetical theory of natural selection*, Clarendon Press

14 Pal, C. and Hurst, L.D. (2003) Evidence for co-evolution of gene order and recombination rate. *Nat. Genet.* 33, 392–395

15 Nei, M. (2003) Genome evolution: let's stick together. *Heredity* 90, 411–412

16 Lawrence, J.G. (2003) Gene organization: selection, selfishness, and serendipity. *Annu. Rev. Microbiol.* 57, 419–440

17 Charlesworth, D. and Charlesworth, B. (1975) Theoretical genetics of Batesian mimicry III. Evolution of dominance. *J. Theor. Biol.* 55, 325–337

18 Kimura, K. (1956) A model of a genetic system which leads to closer linkage by natural selection. *Evolution* 10, 278–287

19 Akhunov, E.D. *et al.* (2003) Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proc. Natl. Acad. Sci. U. S. A.* 100, 10836–10841

20 Teichmann, S.A. and Veitia, R.A. (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics* 167, 2121–2125

21 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113

22 Poyatos, J.F. and Hurst, L.D. (2004) How biologically relevant are interaction-based modules in protein networks? *Genome Biol.* 5, R93

23 Gerton, J.L. *et al.* (2000) Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11383–11390

24 Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197

25 Hershberg, R. *et al.* (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet.* 21, 138–142

26 Barton, N.H. (1995) A general model for the evolution of recombination. *Genet. Res.* 65, 123–145

27 Kondrashov, A.S. (1988) Deleterious mutations and the evolution of sexual reproduction. *Nature* 336, 435–440

28 Hurst, L.D. *et al.* (2002) Natural selection promotes the conservation of linkage of co-expressed genes. *Trends Genet.* 18, 604–606

29 Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12123–12128

30 Blatt, M. *et al.* (1996) Superparamagnetic clustering of data. *Phys. Rev. Lett.* 76, 3251–3254

0168-9525/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2006.06.003