Natural selection promotes the conservation of linkage of co-expressed genes

Laurence D. Hurst, Elizabeth J. B. Williams and Csaba Pál

Although there is increasing evidence that eukaryotic gene order is not always random, there is no evidence that putatively favourable gene arrangements are preserved by selection more than expected by chance. In yeast (Saccharomyces cerevisiae), for example, co-expressed genes tend to be linked, but whether such gene pairs tend to remain linked more often than expected under null neutral expectations is not known. We show using gene pairs in the S. cerevisiae–Candida albicans comparison that highly co-expressed gene pairs are conserved as pairs at about twice the average rate. However, co-expressed genes also tend to be in close physical proximity and, as expected from a null neutral model, genes (be they co-expressed or not) that are physically close together tend to be retained more often. This physical proximity, however, only accounts for a small proportion of the enhanced degree of conservation of co-expressed gene pairs. These results demonstrate that purely neutralist models of gene order evolution are not realistic.

Published online: 01 November 2002

Much current data suggests that the ‘randomly arranged beans on a string’ model of eukaryotic genomes is not adequate. Not only are certain sorts of genes especially prevalent on the X chromosome[1,2], but in humans [3], flies [4], yeast [5] and worm [6] genes of similar expression profile tend to be clustered. In striking contrast, there is very little evidence to suggest that any putatively adaptive clusters remain conserved more often than expected of any random set of genes, with obvious exceptions such as Hox clusters [7]. Based on a limited sample it has, however, been suggested that co-expressed genes in yeast might be conserved at a higher rate than expected [8], although a broad-scale analysis failed to show that gene orientation (a putative covariate of co-expression), was biased in conserved gene pairs [9].

To address this issue we assembled a dataset of S. cerevisiae gene pairs (i.e. nearest neighbours) for which we could define the orthologue for both genes in Candida (1850 pairs). Orthology was determined using reciprocal best hits in BLAST analysis, as previously described [10]. Chromosomal location in yeast was derived from accession numbers NC_001133–48. Protein sequence and location data for C. albicans was obtained from the Stanford DNA Sequencing and Technology Center website at http://www-sequence.stanford.edu/group/candida/index.html; contig version 6.

Of the 1850 yeast gene pairs with Candida orthologues, we eliminated those that were pairs of tandem duplicates (as defined by pairwise BLAST score E < 10^-2), those that were overlapping or with no space between the genes and those for which we could not define the extent of co-expression between neighbouring genes. This left a total of 1817 gene pairs. The dataset includes 166 pairs in yeast that remain as nearest neighbours in Candida. These we consider to be the gene pairs with conserved linkage. The overall proportion conserved (9%) is low, but this is more a measure of the long times since common ancestry (~200 million years) than an indication of the presence or absence of selection. Indeed, comparison can be made with the evolution of codon usage bias: in yeast, highly expressed genes show strong codon usage bias indicative of selection acting on ‘silent’ point mutations, but in comparisons of these genes with orthologues in Candida, the silent site substitution rate is very high and so close to saturation as to be all but uninterpretable.

To establish whether co-expression is important for retention of linkage, we need to define the extent of co-expression. We took the expression profiles from the microarray data compiled by the Eisen lab (http://rana.lbl.gov/EisenData.htm) and, using normalized data, for each linked pair calculated the Pearson correlation coefficient (r) between the two genes, a measure of their degree of co-expression. If co-expression were important in the retention of a gene pair, then we would expect that as the degree of co-expression goes up, so would the probability of conservation of linkage. However, we have no reason to suppose that this is necessarily a gradual effect. For most gene pairs, the r values simply represent random noise: a small positive value for r should not be taken as evidence of more co-ordinated expression than an equally small negative value. Only when their value is especially high do we suspect some functionally significantly co-ordination in the regulation of the two genes.

Therefore, to provide an indication of whether co-expression is important, we performed a sliding-window analysis of gene pairs organized by the ranked r value, calculating mean r, the proportion conserved and mean intergenic spacer. As can be seen in Fig. 1, at high values of mean r (highly co-expressed genes), the proportion conserved does indeed greatly exceed null expectations. This provides the first whole-genome analysis to indicate that co-expressed genes are conserved more than expected by chance. As expected then, the genes pairs that are conserved have higher r values (i.e. are more likely to be co-expressed) than those that are not conserved (Mann–Whitney U test, P = 0.01).

There is, nonetheless, a difficulty with the interpretation of the above result. Examination of Fig. 1 also indicates that as the mean r increases, the mean intergenic distance decreases. The excess conservation of co-expressed gene pairs might then trivially be explained as a consequence of a null neutral evolution of gene order. The simplest explanation for the conservation of linkage is that gene order re-arrangements (e.g. inversions) occur at random locations, that they are tolerated only if they disrupt intergenic spacer and that all such tolerated re-arrangements are without selective consequences. The tolerated ones then could spread by drift (i.e. neutral evolution). Gene pairs with small intergenic spacer should then be expected to be conserved as nearest neighbours more often. Indeed, as predicted, the genes in conserved pairs are closer together than those in the non-conserved pairs (mean intergenic spacer of unconserved
These were generated by random sampling of 200 of the 1817 genes without replacement, 10,000 times. The proportion of gene pairs conserved and mean intergene spacer length as a function of the mean degree of co-expression ($r$ value) for groupings of 200 genes. All the 1817 genes were ranked by $r$ value (rank of 1 = most highly co-expressed gene pair). We then examined the first 200 genes (ranks 1–200) and determined mean $r$, mean intergene spacer and the proportion conserved. The right-most data on the two plots represents the data from this group. We then moved the genes ranked 2 to 201, then 3 to 202 etc., and in each calculated the same parameters. In each plot, the motion of conserved pairs is apparent asymptote. This is probably a conservative definition, but as it is our intention to ask whether we can detect a signal of selection, using the uppermost 14% of genes (in terms of co-expression) is adequate. In this set, 42 are conserved (16.8%) as opposed to 124 of the remaining 1567 (8%) (G test of independence: $G_1 = 13.91, P = 0.001, P < 0.00005$ from 50,000 randomizations). These co-expressed genes have significantly smaller intergene spacer (median intergene spacer co-expressed pairs = 446.5 bp ±36.2; mean intergene spacer of remaining pairs = 502.2 bp ±16.4: Mann–Whitney U test; $P = 0.03$).

To examine whether this higher degree of conservation was due to the reduced intergene spacer size, we split the data into non-overlapping groupings of approximately equal intergene spacer (1–200 bp, 201–400 bp, etc.) and for each sub-group we determined the number of co-expressed genes that are conserved. In each sub-group, the expected number is the number of co-expressed gene pairs (conserved or not conserved) multiplied by the proportion of genes conserved in the dataset as a whole. The blue line indicates the data for those that are not co-expressed ($n = 1124$). The figure is a sliding-window plot of the data ranked by intergene spacer size using 200 gene pair samples and a jump of one gene pair between windows, as in Fig. 1. The ‘non co-expressed’ set was deduced from a $P$ value for the significance of the observed $r$ value between two genes derived by randomization. Those with $P > 0.05$ were assumed to not be co-expressed. This provides a conservative sample of non-co-expressed genes as, owing to multiple sampling, there will be numerous gene pairs that show spurious significance at the 5% level, but are excluded from this dataset.

We define the 250 genes with the highest $r$ values as being co-expressed. This corresponds approximately to those samples in Fig. 1 with mean $r > 0.52$ and represents the approximate position at which the proportion conserved hits the apparent asymptote. This is probably a conservative definition, but as it is our intention to ask whether we can detect a signal of selection, using the uppermost 14% of genes (in terms of co-expression) is adequate. In this set, 42 are conserved (16.8%) as opposed to 124 of the remaining 1567 (8%) (G test of independence: $G_1 = 13.91, P = 0.001, P < 0.00005$ from 50,000 randomizations). These co-expressed genes have significantly smaller intergene spacer (median intergene spacer co-expressed pairs = 446.5 bp ±36.2; mean intergene spacer of remaining pairs = 502.2 bp ±16.4: Mann–Whitney U test; $P = 0.03$).

To examine whether this higher degree of conservation was due to the reduced intergene spacer size, we split the data into non-overlapping groupings of approximately equal intergene spacer (1–200 bp, 201–400 bp, etc.) and for each sub-group we determined the number of co-expressed genes that are conserved. In each sub-group, the expected number is the number of co-expressed gene pairs (conserved or not conserved) multiplied by the proportion of genes conserved in the dataset as a whole. The blue line indicates the data for those that are not co-expressed ($n = 1124$). The figure is a sliding-window plot of the data ranked by intergene spacer size using 200 gene pair samples and a jump of one gene pair between windows, as in Fig. 1. The ‘non co-expressed’ set was deduced from a $P$ value for the significance of the observed $r$ value between two genes derived by randomization. Those with $P > 0.05$ were assumed to not be co-expressed. This provides a conservative sample of non-co-expressed genes as, owing to multiple sampling, there will be numerous gene pairs that show spurious significance at the 5% level, but are excluded from this dataset.

We define the 250 genes with the highest $r$ values as being co-expressed. This corresponds approximately to those samples in Fig. 1 with mean $r > 0.52$ and represents the approximate position at which the proportion conserved hits the apparent asymptote. This is probably a conservative definition, but as it is our intention to ask whether we can detect a signal of selection, using the uppermost 14% of genes (in terms of co-expression) is adequate. In this set, 42 are conserved (16.8%) as opposed to 124 of the remaining 1567 (8%) (G test of independence: $G_1 = 13.91, P = 0.001, P < 0.00005$ from 50,000 randomizations). These co-expressed genes have significantly smaller intergene spacer (median intergene spacer co-expressed pairs = 446.5 bp ±36.2; mean intergene spacer of remaining pairs = 502.2 bp ±16.4: Mann–Whitney U test; $P = 0.03$).

To examine whether this higher degree of conservation was due to the reduced intergene spacer size, we split the data into non-overlapping groupings of approximately equal intergene spacer (1–200 bp, 201–400 bp, etc.) and for each sub-group we determined the number of co-expressed genes that are conserved. In each sub-group, the expected number is the number of co-expressed gene pairs (conserved or not conserved) multiplied by the proportion of genes conserved in the dataset as a whole. The blue line indicates the data for those that are not co-expressed ($n = 1124$). The figure is a sliding-window plot of the data ranked by intergene spacer size using 200 gene pair samples and a jump of one gene pair between windows, as in Fig. 1. The ‘non co-expressed’ set was deduced from a $P$ value for the significance of the observed $r$ value between two genes derived by randomization. Those with $P > 0.05$ were assumed to not be co-expressed. This provides a conservative sample of non-co-expressed genes as, owing to multiple sampling, there will be numerous gene pairs that show spurious significance at the 5% level, but are excluded from this dataset.

We define the 250 genes with the highest $r$ values as being co-expressed. This corresponds approximately to those samples in Fig. 1 with mean $r > 0.52$ and represents the approximate position at which the proportion conserved hits the apparent asymptote. This is probably a conservative definition, but as it is our intention to ask whether we can detect a signal of selection, using the uppermost 14% of genes (in terms of co-expression) is adequate. In this set, 42 are conserved (16.8%) as opposed to 124 of the remaining 1567 (8%) (G test of independence: $G_1 = 13.91, P = 0.001, P < 0.00005$ from 50,000 randomizations). These co-expressed genes have significantly smaller intergene spacer (median intergene spacer co-expressed pairs = 446.5 bp ±36.2; mean intergene spacer of remaining pairs = 502.2 bp ±16.4: Mann–Whitney U test; $P = 0.03$).

To examine whether this higher degree of conservation was due to the reduced intergene spacer size, we split the data into non-overlapping groupings of approximately equal intergene spacer (1–200 bp, 201–400 bp, etc.) and for each sub-group we determined the number of co-expressed genes that are conserved. In each sub-group, the expected number is the number of co-expressed gene pairs (conserved or not conserved) multiplied by the proportion of genes conserved in the dataset as a whole. The blue line indicates the data for those that are not co-expressed ($n = 1124$). The figure is a sliding-window plot of the data ranked by intergene spacer size using 200 gene pair samples and a jump of one gene pair between windows, as in Fig. 1. The ‘non co-expressed’ set was deduced from a $P$ value for the significance of the observed $r$ value between two genes derived by randomization. Those with $P > 0.05$ were assumed to not be co-expressed. This provides a conservative sample of non-co-expressed genes as, owing to multiple sampling, there will be numerous gene pairs that show spurious significance at the 5% level, but are excluded from this dataset.

We define the 250 genes with the highest $r$ values as being co-expressed. This corresponds approximately to those samples in Fig. 1 with mean $r > 0.52$ and represents the approximate position at which the proportion conserved hits the apparent asymptote. This is probably a conservative definition, but as it is our intention to ask whether we can detect a signal of selection, using the uppermost 14% of genes (in terms of co-expression) is adequate. In this set, 42 are conserved (16.8%) as opposed to 124 of the remaining 1567 (8%) (G test of independence: $G_1 = 13.91, P = 0.001, P < 0.00005$ from 50,000 randomizations). These co-expressed genes have significantly smaller intergene spacer (median intergene spacer co-expressed pairs = 446.5 bp ±36.2; mean intergene spacer of remaining pairs = 502.2 bp ±16.4: Mann–Whitney U test; $P = 0.03$).

To examine whether this higher degree of conservation was due to the reduced intergene spacer size, we split the data into non-overlapping groupings of approximately equal intergene spacer (1–200 bp, 201–400 bp, etc.) and for each sub-group we determined the number of co-expressed genes that are conserved. In each sub-group, the expected number is the number of co-expressed gene pairs (conserved or not conserved) multiplied by the proportion of genes conserved in the dataset as a whole. The blue line indicates the data for those that are not co-expressed ($n = 1124$). The figure is a sliding-window plot of the data ranked by intergene spacer size using 200 gene pair samples and a jump of one gene pair between windows, as in Fig. 1. The ‘non co-expressed’ set was deduced from a $P$ value for the significance of the observed $r$ value between two genes derived by randomization. Those with $P > 0.05$ were assumed to not be co-expressed. This provides a conservative sample of non-co-expressed genes as, owing to multiple sampling, there will be numerous gene pairs that show spurious significance at the 5% level, but are excluded from this dataset.
in the co-expressed class, whereas the apparent enrichment of conservation of gene pairs is highly significant.

Prior evidence suggests that divergently oriented genes (=→) are especially likely to belong to a single regulatory unit [11]. Within the highly co-expressed group, the genes in divergent orientation are indeed more common than expected from their overall frequency (we expect 85, but observe 115). Contrary to previous suggestions [11], we find a dearth of gene pairs in which both genes are in the same orientation (87 observed, 115 expected). Overall, there is a significant difference in the proportion of types in different orientations in the highly co-expressed class compared with the distribution in the dataset as a whole (x^2 = 13.88, 2, P < 0.001).

Of the 42 pairs that are conserved within the highly co-expressed class, 19 (45%) are in the divergent orientation in yeast, approximately double their frequency within the dataset as a whole (G test of independence, P < 0.01), and higher than their frequency within the co-expressed class, although not significantly so (G test of independence, P > 0.05). Although the above results suggest that divergent orientation is important for co-regulation and for conservation of pairs, we do not find that the divergent genes retain their orientation at an especially high rate. Of the 19, 14 (74%) have the same orientation in Candida, which compares with 62% of conserved pairs that have the same direction in both species (i.e. 103 out of the sample of 166). Nonetheless, our findings are consistent with the observation [12] that between S. cerevisiae and Candida albicans, divergently transcribed gene-pairs that are conserved in evolution have a higher probability of being co-regulated than divergently transcribed gene pairs that are disrupted in evolution.

We conclude that, consistent with the null neutral model, gene pairs that have small intergene spacer are the most likely to be conserved. This result emphasizes the need to control for the length of intergene spacer size. The x axis of Fig. 3 shows the upper limit to the subgroup size (i.e. 200 represents genes with intergene spacer less than or equal to 200, 400, indicate 201 to 400 etc.).

References

Acknowledgements
We should like to thank Martijn Huynen and an anonymous referee for comments on an earlier version of the manuscript. C.P. is funded by a Royal Society/Nato visiting fellowship and L.D.H. by B.B.S.R.C.

As a result, we do not find that the divergent genes retain their orientation at an especially high rate. Of the 19, 14 (74%) have the same orientation in Candida, which compares with 62% of conserved pairs that have the same direction in both species (i.e. 103 out of the sample of 166). Nonetheless, our findings are consistent with the observation [12] that between S. cerevisiae and Candida albicans, divergently transcribed gene-pairs that are conserved in evolution have a higher probability of being co-regulated than divergently transcribed gene pairs that are disrupted in evolution.

We conclude that, consistent with the null neutral model, gene pairs that have small intergene spacer are the most likely to be conserved. This result emphasizes the need to control for the length of intergene spacer size. The x axis of Fig. 3 shows the upper limit to the subgroup size (i.e. 200 represents genes with intergene spacer less than or equal to 200, 400, indicate 201 to 400 etc.).

Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice

Klaas Vandepoele, Cedric Simillion and Yves Van de Peer

Genome analysis shows that large-scale gene duplications have occurred in fungi, animals and plants, creating genomic regions that show similarity in gene content and order. However, the high frequency of gene loss reduces colinearity resulting in duplicated regions that, in the extreme, no longer share homologous genes. Here, we show that by comparison with an appropriate second genome, such paralogous regions can still be identified.

Published online: 30 October 2002