Chance and necessity in chromosomal gene distributions

Rutger Hermsen,[*] Pieter Rein ten Wolde[*] and Sarah Teichmann[†]

March 8, 2008

Abstract

By analyzing the spacing of genes on chromosomes, we find that transcriptional and RNA-processing regulatory sequences outside coding regions leave footprints on the distribution of intergenic distances. Using analogies between genes on chromosomes and one-dimensional gases, we constructed a statistical null model. We have used this to estimate typical upstream and downstream regulatory sequence sizes in various species. Deviations from this model reveal bi-directional transcriptional regulatory regions in *S. cerevisiae* and bi-directional terminators in *E. coli*.

1 Probability distributions of intergenic distances

The probability distributions of intergenic distances are shaped by stochastic processes, such as insertions, deletions, inversions and duplications, and by natural selection. Although the former tends to randomize the distribution, the latter introduces biases if there are functional reasons for genes to be spaced in a particular way [1]. Here we compare data of both *Escherichia coli* and different fungal species to a statistical mechanics model to study which features can be explained by random processes only, and which require an explanation in terms of functionality.

2 The Constant-Force model

The Constant-Force (CF) model is based on two observations (see Fig. 1). First, open reading frames (ORFs) usually do not overlap, even if their density is high. For example, if the genes of *Saccharomyces cervisiae* (budding yeast) were randomly distributed, 78% of the ORFs would overlap with another ORF, whereas in reality, only 9% do.

Second, ORFs are rarely very close together (e.g. see Fig. 2a for *S. cerevisiae*). We hypothesize that this is caused by functional sequences directly upstream and downstream

[*]FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ, Amsterdam, The Netherlands
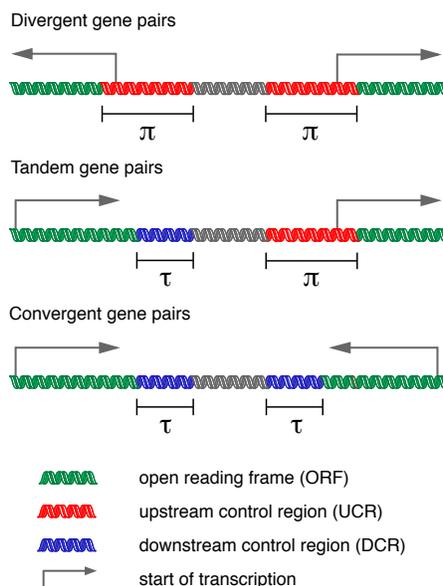[†]MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK CB2 2QH

Figure 1: The Constant-Force model. This model assumes that a 5'-UTR, a basal promoter and a *cis*-regulatory region are present upstream of every ORF. We call this the upstream control region (UCR), and assume it has a fixed size π . Downstream of each ORF, the 3'-UTR and possibly a transcriptional terminator and RNA processing signals are present, to which we jointly refer as the downstream control region (DCR), assumed to have length τ. The figure shows that ORFs neighboring on the DNA can have three mutual orientations: divergent (D), tandem (T) or convergent (C). This also leads to three kinds of intergenic regions: D regions contain two UCRs, while T regions contain one UCR and one DCR, and C regions have two DCRs. In *S. cerevisiae*, the frequencies of D, T and C regions are 26.3%, 48.3% and 25.4% respectively, which is close to the random proportions 1:2:1. This holds for most fungi. In *E. coli*, T regions are more frequent due to the organization of its genes in operons (17.5%, 66.7% and 15,7%).
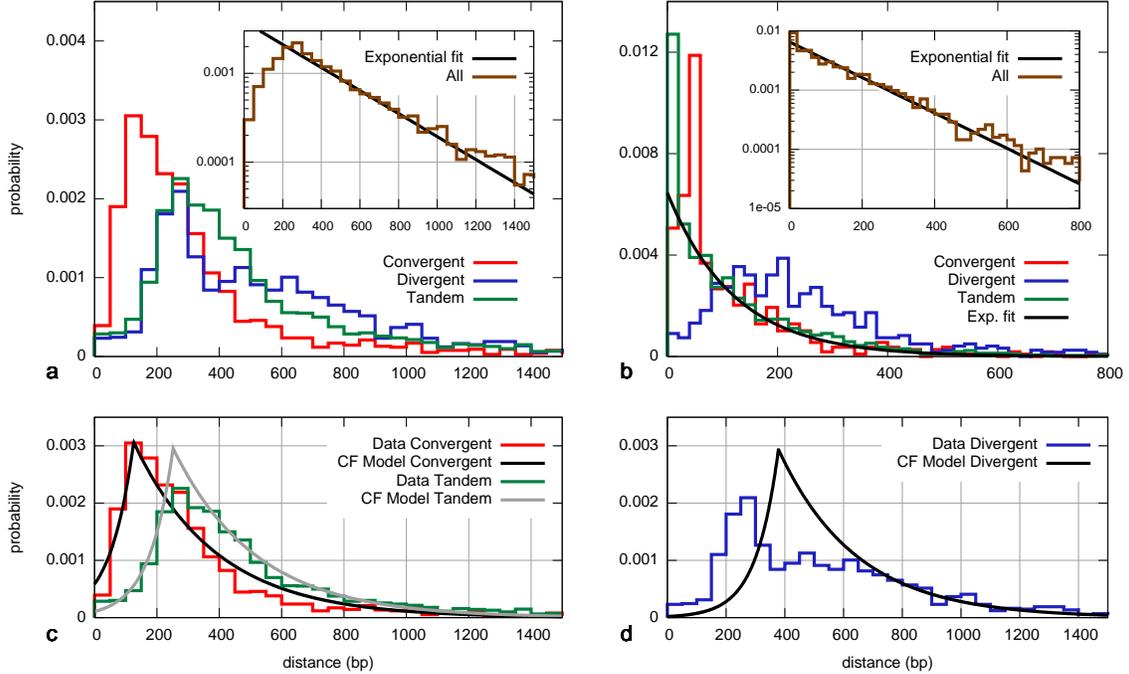
Figure 2: Probability distributions of intergenic distances in *S. cerevisiae* and *E. coli*. (a) Probability distributions of intergenic regions in *S. cerevisiae*. The distributions of distances between convergent (C), divergent (D) and tandem (T) gene pairs. C intergenic regions are, on average, shorter than T regions; the D regions are longest. Note also that the divergent distribution has a bimodal shape, with a peak at $n \approx 275$ bp and one at $n \approx 500$ bp. Inset: the distribution of all intergenic regions is exponential for distances larger than 300 bp (scale parameter: 335 bp), but has a "dip" at shorter distances. This dip, we argue, is a footprint of UCRs and DCRs. (b) As panel (a), but for *E. coli*. The T distribution in the main plot is exponential, except for an accumulation in the first bin. This accumulation is the result of intergenic regions inside operons, which are not separated by control regions and therefore can be arbitrarily close together [2]. The C distribution is also exponential, except for a peak at 20–60 bp, where *S. cerevisiae* has a dip instead. We predict that this peak is the result of bi-directional terminator sequences. The inset again shows that the distribution of all intergenic regions is largely exponential (scale parameter: 145 bp). (c) Simultaneous fit of the Constant Force (CF) model to the C and T distributions of *S. cerevisiae*. The model fits the data surprisingly well. (d) The D distribution of *S. cerevisiae* and the expected distribution according to the CF model. Clearly, the bi-modal shape of the data is not consistent with the CF model. We predict that the set of divergent intergenic regions in *S. cerevisiae* consists of two subpopulations: those containing two independent *cis*-regulatory regions, responsible for the second peak, and those containing one bi-directional *cis*-regulatory region.

3

of the ORFs, which we call upstream and downstream control regions (UCRs and DCRs). UCRs include basal promoters, *cis*-regulatory regions and 5' untranslated regions (UTRs); DCRs consist of 3'-UTRs, transcriptional terminators and RNA-processing signals. If ORFs approach each other closely, these regions need either to overlap or to be very short, which makes such configurations less likely. To test this, we divide the intergenic regions into three subsets, called tandem (T), convergent (C) and divergent (D). (See Fig. 1.) Intergenic regions in subset T should contain one DCR and one UCR, whereas C and D intergenic regions contain two DCRs and two UCRs respectively. As UCR sequences are generally longer than DCRs, we expect that D regions are on average longer than T regions and C regions are shortest, which is indeed the case (see Fig. 2a and supplementary Fig. 3).

These observations inspire the following model. We assume that all ORF configurations are equally probable, except for the following constraints: (i) ORFs do not overlap; (ii) UCRs and DCRs can overlap with each other or with ORFs, but every overlapping base pair (bp) in a particular configuration makes this configuration a factor q less probable. For simplicity, we assume that in a given organism all UCRs and DCRs have a fixed length, π and τ respectively.

This model is equivalent to a one-dimensional system of hard particles with a finite-ranged, repulsive, *constant-force* interaction. Tandem, convergent and divergent ORF pairs interact at a range $\pi + \tau$, 2τ and 2π , respectively. This mapping enables us to use the formalism of statistical physics to compute the probability distributions corresponding to this model analytically (see the supplementary material).

3 The CF model fits the C and T distribution of *S. cerevisiae*

The CF model fits the distributions of *S. cerevisiae* convergent and tandem intergenic distances remarkably well (see Fig. 2c). The fit parameters are $\tau = 61$ bp for DCRs and $\pi = 196$ bp for UCRs. These numbers provide a course estimate of the space required for the transcriptional and translational regulatory signals and RNA processing in *S. cerevisiae*.

Our UCR length prediction of 196 bp is in excellent agreement with the distribution of transcription-factor-binding sites near *S. cerevisiae* start codons, which has its peak at 100-200bp from the start codon [3]. Our DCR prediction of 61 bp is supported by bioinformatics analyses of *S. cerevisiae* 3'-RNA processing signals, which show that the majority of these sequences is within 20–90 bp of the stop codon [4]. However, Graber *et al.* predict longer 3'-UTRs [5] and recent experiments show that the median of 3'-UTRs lengths is ≈ 91 bp [6], which suggests that our DCR estimate is on the low side. (In the supplementary material, we show that more refined models can provide quantitative agreement).

4

4 Typical UCR and DCR sizes for other fungi

Name of organism	UCR length/bp	DCR length/bp	q
S. cerevisiae	196 ± 4	61 ± 1	0.985 ± 0.001
C. glabrata	296 ± 4	66 ± 2	0.983 ± 0.001
K. lactis	295 ± 5	38 ± 2	0.987 ± 0.001
D. hansenii	141 ± 4	28 ± 2	0.976 ± 0.001

Table 1: Estimates for the UCR and DCR sizes for various fungal species. The errors given in the table are the uncertainties of the fit parameters and as such should not be interpreted as variances of these quantities in the genome.

We repeated this approach to estimate the UCR and DCR lengths for three additional fungi using only the ORF coordinates as input. (See Table 1 and Fig. 4 in the supplementary material). We found that UCRs are consistently longer than DCRs. The UCR and DCR lengths seem to vary independently of each other, and no dependence on gene density is apparent. Recently, it has been shown that the distribution of rigid DNA in *cis*-regulatory regions of fungi correlates with the position of transcription-factor-binding sites [7]. Our estimates for the UCR lengths correlate well with the position of rigid DNA in these fungi.

5 *S. cerevisiae* contains many bi-directional UCRs

We now turn to the spacing of divergent pairs in budding yeast. Interestingly, the corresponding distribution has a bimodal shape (see Fig. 2d) that is even more pronounced in other fungi (Fig. 3 in the supplementary material). The first, narrow peak is centered on bp; the second peak is broader and is maximal around bp. This shape is not consistent with the CF model.

Apparently, many divergent intergenic regions are very short: 29% are < 300 bp. Because few (10%, 280 out of 2801) tandem intergenic regions, containing only one UCR, are < 200 bp, it seems unlikely that two independent UCRs could fit in divergent intergenic regions with a length of the order of 275 bp. Hence we propose that the set of divergent gene pairs is composed of two sub-populations.

The first population, corresponding to the second peak, consists of pairs of genes that are regulated independently. The other sub-population consists of gene pairs that share a bi-directional *cis*-regulatory region, that is, a regulatory region containing elements such as transcription factor binding sites that regulate the expression of both flanking genes. Such a coupling could force genes to preserve their proximity, thus causing the deviation from the CF model. While bi-directional *cis*-regulatory regions are ubiquitous in *E. coli* [1], only a few bi-directional UCRs have been reported in *S. cerevisiae* [8, 9, 10, 11]. Based on

Fig. 2d, we predict that about 30% (426 out of 1471) of the divergent pairs are regulated by a shared *cis*-regulatory region. (We list the best candidates in the Supplement.)

If this is true, then one would expect co-expressed divergent pairs to be overrepresented in the first peak rather than the second. This is indeed the case for positively correlated pairs ($p < 0.002$; see Supplement, section D.1). Negatively correlated pairs are typically not in the first peak. This contrasts with bi-directional UCRs in bacteria, in which dual regulators often act as a repressor for one of the genes and as an activator for the other, resulting in anti-correlated expression patterns.

We also used Gene Ontology (GO) annotations [12] to test whether the divergent neighbors in the first peak are more often functionally related than those in the second peak. Adopting the method of reference [13] to quantify the similarity between GO terms, we indeed found this to be the case ($p < 9 \times 10^{-4}$ for biological process, $p < 5 \times 10^{-5}$ for cellular component; see Supplement, section D.2).

6 *E. coli* has many bi-directional terminators

As mentioned above, bi-directional promoters are well-characterized in *E. coli*. We now show that the distribution of convergent gene pairs in *E. coli* provides evidence for bi-directional transcriptional terminators, which are much less well described.

In accordance with the CF model, the C distribution has an exponential signature (see Fig. 2b). Convergent intergenic regions are expected to contain two DCRs. Given the typical size of Rho-independent terminators ($\approx$ 40 bp) the CF model predicts a dip at short distances (< 80 bp). Instead, there is a significant excess of intergenic regions of size 20 to 60 bp ($p = 10^{-13}$; see Supplement, section D.3). It is unlikely that two terminators would fit into such short intergenic regions.

Rho-independent terminator sequences function by stem-loop formation of the RNA transcript, and hence are largely palindromic. As the complementary strand of a palindromic sequence is a palindrome too, some terminators can function bi-directionally. Indeed, a few bi-directional terminators have been identified experimentally [14]. Moreover, Lesnik *et al.* used an algorithm called RNAMotif to identify putative terminators and predicted that many of them could function bi-directionally [2]. Given that most terminators in *E. coli* start within 60 bp downstream of their ORF [2], genes sharing a bi-directional terminator should usually be close together; this suggests that the peak in the distribution at short distances is caused by bi-directional terminators.

To explain the data, at least 86 bi-directional terminators should be present; this would imply that as many as 23% of the operons use a bi-directional terminator (see Supplement, section D.3).

We tested this, using the data of Lesnik *et al.*[1]. Indeed, putative terminators that RNAMotif classifies as bi-directional have a tendency to occur in short, convergent inter-

---

[1]Although no statistical test is presented, a similar conclusion is reached in ref. [15]

genic regions, corroborating our hypothesis ($p < 0.0003$, see Supplement, sections D.4 and D.5) [16].

7 Concluding remarks

The largest limitation of the current model is the assumption that the UCRs and DCRs have fixed sizes. Especially in higher eukaryotes, UCR and DCR lengths often have a high variance; in these cases, it is necessary to include this in the model. In the Supplement we show that this can be done and how more realistic potentials can be chosen.

That being said, the simple CF model describes many universal characteristics of the gene spacing. It not only quantitatively describes the exponential decay at large distances, but also the repulsion at short distances due to UCRs and DCRs in prokaryotes and eukaryotes alike. In *E. coli* and fungi, this repulsion provides information about the typical length of UCRs and DCRs using only the ORFs coordinates as input. The model can also serve as a null model for the spacing of genes: deviations from it lead to meaningful predictions about the presence of operons, bi-directional promoters or terminators.

8 Acknowledgements

Supplementary Text

A   The Constant Force Model

In this section we provide additional information about the Constant-Force (CF) model.

A.1   Assumptions

As we explained in the main text, the CF model is based on three assumptions. First, we assume that ORFs cannot overlap. Second, in a given organism, upstream control regions (UCRs) and downstream control regions (DCRs) have a fixed size (π and τ respectively). Third, we assume that these control regions can overlap with each other and with nearby ORFs, but that such overlaps are not likely. More precisely, we assume that *whenever a base pair from such a region overlaps with another functional region, be it an ORF, a UCR or a DCR, it makes that particular configuration a factor q less probable.* For simplicity, we make no distinction between the different kinds of overlap.

A useful analogy can be drawn with a physical system. The proposed model is formally equivalent to a one-dimensional system of hard particles with finite-ranged repulsive interactions. The interaction determined by our assumptions is an interaction with a *constant force.* The range of the interaction depends on the mutual orientation of the neighboring genes. Divergent gene pairs are separated by two UCRs and therefore start interacting at a distance 2π; convergent pairs have two DCRs in their intergenic region and therefore have an interaction range of 2τ, and intergenic regions between tandem pairs contain one DCR and one UCR (interaction range $\pi + \tau$). This analogy allows us to use the formalism of statistical physics to compute the probability distribution of the intergenic distances for this model analytically; the complete derivation follows below.

A.2   Derivation of the distance distributions: CF interaction with fixed range

In the CF model described above, the interaction range of the particles depends on their mutual orientation (convergent, divergent or tandem). We first derive the distance distribution for a slightly simpler system, in which the interaction range does not depend on the orientation.

We consider a one-dimensional space (representing the chromosome) of length L' containing $N-1$ particles (representing ORFs). We choose to describe the system in the micro-canonical ensemble, with fixed total energy E. The state of the system can be described by a vector $\vec{n} = (n_1, n_2, \ldots, n_N)$, where n_i is the length of the ith inter-particle space. The sum of these numbers, $L \equiv \sum_i n_i$, is the total free space in the system. The value of L is fixed and $L \gg 1$. As the particles occupy part of the total space, $L < L'$.

For now we assume that the particles interact with a finite-ranged CF potential $U(n)$, defined as:

$$\frac{U(n)}{kT} = \begin{cases} \epsilon(r - n) & (n < r) \\ 0 & (n \geq r), \end{cases} \tag{1}$$

where r is the range of the interaction, and ϵ is the energy associated with an overlap of one base pair (in units of kT); it is related to q as $\epsilon = -\ln(q)$.

In order to compute the probability distribution of intergenic distances, we divide the system into two subsystems. Subsystem 1 (S1) is a particular, but arbitrary, inter-particle space x, while subsystem 2 (S2) is the rest of the system. We will compute the probability distribution $P(n_x)$ of the length n_x of space x. We define the multiplicity function of subsystem S2, called $\Omega_2(L_2, E_2)$, as the number of states accessible for S2 given the available free length for S2, L_2, and the available energy for S2, E_2. Note that $L_2 = L - n_x$ and $E_2 = E - E_1 = E - U(n_x)$. Then the probability that inter-particle space x has length n_x is proportional to the number of states that are accessible to the rest of the system, S2, given that x has length n_x:

$$P(n_x) \propto \Omega_2\left(L - n_x, E - U(n_x)\right). \tag{2}$$

By definition, the entropy $\sigma_2(L_2, E_2)$ of S2 is the logarithm of $\Omega_2(L_2, E_2)$. Therefore,

$$P(n_x) \propto e^{\sigma_2(L - n_x, E - U(n_x))}. \tag{3}$$

Assuming that n_x is small compared to L and that $U(n_x)$ is small compared to E, we can now expand the entropy as follows:

$$\begin{aligned} \sigma_2(L - n_x, E - U(n_x)) = \quad & \sigma_2(L, E) \\ & - \; n_x \frac{\partial \sigma_2(L, E)}{\partial L} \\ & - \; U(n_x) \frac{\partial \sigma_2(L, E)}{\partial E} \\ & + \; \ldots \end{aligned} \tag{4}$$

Note that by the standard Maxwell relations, $\left(\frac{\partial \sigma_2}{\partial E}\right)_L = 1/kT$ and $\left(\frac{\partial \sigma_2}{\partial L}\right)_E = p/kT$, where T and p are the temperature and the pressure of the system. If N is large, the higher order terms are negligible.

Now we can combine the expansion in equation 4 with equation 3 and obtain:

$$P(n_x) \propto e^{-(n_x p + U(n_x))/kT}. \tag{5}$$

We can calculate this in full using the definition of the potential in equation 1, arriving at:

$$P(n_x) = \begin{cases} c \, e^{-n_x(\lambda - \epsilon)} & (n_x < r) \\ c \, e^{-n_x \lambda + \epsilon r} & (n_x \geq r). \end{cases} \tag{6}$$

9

Here λ is defined as $\lambda \equiv \frac{p}{kT}$. As we picked inter-particle space x arbitrarily, this probability distribution holds for all inter-particle spaces. The number c is a normalization constant. Given r and ϵ, the value of λ is fixed if we impose the mean inter-particle distance:

$$\int n_x P(n_x) dn_x = L/N. \tag{7}$$

Note that, beyond the interaction range, the distribution is exponentially decreasing. Within the interaction range, the distribution is also exponential, but the sign of the exponent depends on the size of ϵ: if the repulsion is strong ($\epsilon > \lambda$), the exponent becomes positive in the interaction range. We also note that if either the range r or the repulsion ϵ is set to zero, the distance distribution simply becomes a single exponential. The resulting model is known as a Tonks gas [17].

A.3 CF interactions with different ranges

In the previous subsection we discussed a CF model in which each particle interacts with its neighbors according to one fixed interaction range. In the relevant case, however, the interaction range depends on the mutual orientation of the particles. The interaction potentials for convergent (C), tandem (T) and divergent (D) pairs can be written as follows:

$$
\begin{aligned}
\frac{U_C(n)}{kT} &= \begin{cases} \epsilon(2\tau - n) & \text{if } n \le 2\tau, \\ 0 & \text{if } n > 2\tau, \end{cases} \\
\frac{U_T(n)}{kT} &= \begin{cases} \epsilon(\tau + \pi - n) & \text{if } n \le \tau + \pi, \\ 0 & \text{if } n > \tau + \pi, \end{cases} \\
\frac{U_D(n)}{kT} &= \begin{cases} \epsilon(2\pi - n) & \text{if } n \le 2\pi, \\ 0 & \text{if } n > 2\pi. \end{cases}
\end{aligned} \tag{8}
$$

It is rather straightforward to adjust the calculations in the previous section to this case. We again divide the system in two parts, S1 and S2, in which S1 consists of one inter-particle region called x, and S2 is the rest of the system. The derivation in the previous section applies without alteration up to equation 5, irrespective of the orientation corresponding to x (that is: D, T or C). Only in the step from equation 5 to equation 6, the difference in the potentials for D, T and C becomes relevant. As a result, the distributions for the D, C and T intergenic regions all have the form of equation 6, except for a different range r, and a different normalization factor c:

$$
\begin{aligned}
P_C(n) &= \begin{cases} c_1 e^{-n(\lambda - \epsilon)} & \text{if } 0 \le n \le 2\tau, \\ c_1 e^{-n\lambda + 2\tau\epsilon} & \text{if } n > 2\tau, \end{cases} \\
P_T(n) &= \begin{cases} c_2 e^{-n(\lambda - \epsilon)} & \text{if } 0 \le n \le \tau + \pi, \\ c_2 e^{-n\lambda + (\tau + \pi)\epsilon} & \text{if } n > \tau + \pi, \end{cases} \\
P_D(n) &= \begin{cases} c_3 e^{-n(\lambda - \epsilon)} & \text{if } 0 \le n \le 2\pi, \\ c_3 e^{-n\lambda + 2\pi\epsilon} & \text{if } n > 2\pi. \end{cases}
\end{aligned} \tag{9}
$$

Here the prefactors c_1, c_2 and c_3 are defined as

$$
\begin{aligned}
c_1 &= \frac{\lambda(\lambda - \epsilon)}{\lambda - e^{2\tau(\epsilon - \lambda)}\epsilon}, \\
c_2 &= \frac{\lambda(\lambda - \epsilon)}{\lambda - e^{(\tau + \pi)(\epsilon - \lambda)}\epsilon}, \\
c_3 &= \frac{\lambda(\lambda - \epsilon)}{\lambda - e^{2\pi(\epsilon - \lambda)}\epsilon}.
\end{aligned}
\tag{10}
$$

We note that the CF model has four parameters: λ, τ, π, and ϵ. However, if we impose the average length of the intergenic regions, this again leads to a constraint that eliminates one of the parameters. As the total system is a mixture of D, C and T intergenic regions in proportions $f_D : f_C : f_T$ (in most genomes roughly 1:1:2), this constraint becomes:

$$
\int n \frac{f_D P_D(n) + f_C P_C(n) + f_T P_T(n)}{f_D + f_C + f_T} dn = L/N.
\tag{11}
$$

We used Monte Carlo simulations to check the validity of these equations and found excellent agreement.

B More detailed models

The CF model is purposely oversimplified. Such simplified models, with few parameters, provide insight into the essential ingredients of the mechanisms studied. At the same time the simplicity of the CF model leads to certain artifacts. Here we show that such artifacts can be alleviated by more detailed models. Below we discuss how one can allow for varying UCR and DCR lengths, and how alternative interaction potentials can be chosen, with distance-dependent forces.

B.1 Polydisperse UCRs and DCRs

The distributions of the CF model have a sharp peak; this is an artifact of our assumption that all UCRs and all DCRs have the same length. We can extend the model to describe systems with varying UCR and DCR lengths.

If UCR and DCR lengths vary, then this results in a varying interaction range r. In general, due to differences in the UCR and DCR lengths, the interaction range obeys probability distributions $p_C(r)$, $p_D(r)$ and $p_T(r)$ for the convergent, divergent and tandem intergenic regions respectively. Then at a given pressure λ the distributions of intergenic distances are given by

$$
P_C(n) = \int_0^\infty p_C(r) P(n|\lambda, r) dr,
\tag{12}
$$

11

$$P_D(n) = \int_0^\infty p_D(r)P(n|\lambda, r)dr,$$

$$P_T(n) = \int_0^\infty p_T(r)P(n|\lambda, r)dr.$$

Here $P(n|\lambda, r)$ is the probability distribution for the length n of an intergenic region, given the interaction range r and the pressure λ; it depends on the the form of the interaction potential. For instance, if the potential is that of the CF model (equation 1), then $P(n|\lambda, r)$ is given by equation 6. Note that we retrieve the original CF model if we insert $p_C(r) = \delta(r - 2\tau)$, $p_D(r) = \delta(r - 2\pi)$ and $p_T(r) = \delta(r - (\tau + \pi))$ in the above integrals.

In the case of *S. cerevisiae* some studies [6, 4, 5, 18] suggest that the distribution of 3'-UTRs is log-normal. We therefore assume that $p_C(r)$ is the distribution of the sum of two numbers drawn independently from a log-normal distribution. A sum of log-normally distributed random variables can be approximated reasonably by another log-normal distribution. We therefore assume that $p_C(r)$ is log-normal as well (with parameters μ and σ). The corresponding fit to the histogram of convergent intergenic distances in *S. cerevisiae* is better than the fit of the CF model and does not show the artifactual sharp peak (see Fig. 3). Nevertheless, the mean of the best-fitting log-normal distribution ($\mu = 4.61$, $\sigma = 0.405$, mean $= e^{\mu + \sigma^2/2} = 109$) is rather close to the estimate resulting from the CF model. ($2\tau = 122$). Below we show that a better agreement with experiment can be obtained if we allow for an alternative interaction potential.
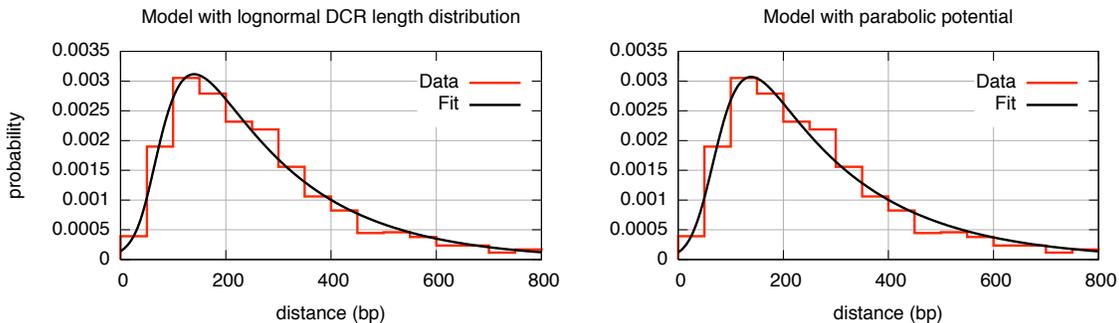

Figure 3: Fits of two more detailed models to the length distribution of convergent intergenic regions in *S. cerevisiae*. The fact that both models allow for nearly perfect fits shows that one needs additional, independent information to distinguish between the various compatible models. Left: CF model with log-normally distributed DCR lengths. Fit parameters for the log-normal distribution: $\mu = 4.61$, $\sigma = 0.405$, $\lambda = 5.25$, $\epsilon = 4.70 \times 10^{-2}$. Right: Log-normally distributed DCR lengths and parabolic potential. Fit parameters: $\mu = 5.01$, $\sigma = 0.314$, $\lambda = 4.95 \times 10^{-3}$, $U_0 = 4.11$.

B.2 Alternative potentials

In the CF model, we used the simple potential defined in equation 1. This potential was convenient because of its simplicity (only one parameter) and its straightforward interpretation. It is, however, possible to generalize our approach to alternative potentials. Equation 5 holds for any finite-ranged potential $U(n)$; this means that equations 5 and 12 can be used to compute the ORF spacing for arbitrary finite-range potentials.

B.3 Yeast DCRs

In the main text, we mentioned that the CF model predictions for the DCR length are on the low side. Not much is known about termination sequences in *S. cerevisiae*, but most of the poly-adenylation signals seem to occur within 70 bp from the stop codon [4]. Estimates for the median 3' UTR length in *S. cerevisiae* range from 80 to about 100 bp[6, 5]. This shows that, although the CF model does predict the qualitative features of the distributions in Yeast, such as the exponential tail of the distribution, in order to get accurate quantitative agreement with the DCR lengths found in recent experiments, the assumptions of the CF model are too crude. Using the above techniques, we can refine the model and get better agreement.

First, recent studies suggest that the 3'-UTRs in Yeast can be approximated by a log-normal distribution; we therefore now choose $p_C(r)$ to be log-normal (with parameters μ and σ). Second, recent experiments strongly suggest that many 3'-UTRs are long and that they often overlap considerably [6, 18]; nevertheless, the ORFs hardly ever get closer together than 120 bp. This suggests a model in which the force is not constant; instead, the repulsion seems to be high at short distances, but low at longer distances. One way to model this is to use the following quadratic potential instead of equation 1:

$$\frac{U(n)}{kT} = \begin{cases} U_0(1 - \frac{n}{r})^2 & (n < r) \\ 0 & (n \geq r). \end{cases} \tag{13}$$

The fit of this model (with μ, σ, U_0 and λ as parameters, but a given mean distance) to the convergent data is excellent (see Fig. 3); also, the resulting log-normal distribution for $p_C(r)$ has a mean $e^{\mu+\sigma^2/2} = 160$, which leads to a mean DCR length of about 80 bps. This is good agreement with the experimental results.

In the study of Van Helden *et al* [4], poly-adenylation signals were found at about 35 bp and 55 bp downstream of the stop codon of ORFs. It is tempting to speculate that these sequences are responsible for the strong repulsion starting at a distance of about 120 bp in convergent intergenic regions.

B.4 Higher eukaryotes

In Fig. 4 and S5 the intergenic distance distributions for various different organisms are shown. Strikingly, the simple CF model can very well describe the qualitative features of all

these model organisms, such as the exponential tail of the distributions and the dependence of the distributions on orientation.

In complex, multicellular eukaryotes, control regions typically are very long and exhibit a high variance(see e.g. [19]). As the lengths and variances increase, the assumptions of the constant force model become less justified. Above we have shown that the CF model can be extended to incorporate alternative potentials and polydisperse interaction ranges. This allows us to produce excellent fits to the data for all organisms. Nevertheless, when it comes to *predicting* the length distributions of UCRs and DCRs for higher organisms, the results depend too sensitively on the choice of the potential to produce meaningful predictions. Therefore we refrain from using the fit parameters for *D. melanogaster*, *A. thaliana*, *C. elegans* and *P. falciparum* as predictions for the DCR and UCR lengths

C Supplementary figures S1 and S2 and fitting procedure

Fig. 4 shows the distributions of intergenic distances for four different fungi and four additional eukaryotes, broken down into three different subsets (convergent, tandem, divergent). The C and T distributions are also displayed in Fig. 5 in log-linear scale, combined with fits of the CF model. In case of the fungi, we used these fits to estimate UCR and DCR sizes in these species. The fit parameters are given in Table 1 in the main text.

We used the maximum likelihood method to fit our model to the data and to determine the errors in the fit parameters. For a given set of observed intergenic distances ($\{n_C\}$ and $\{n_T\}$ for convergent and tandem pairs, respectively), Bayes' rule states that the likelihood of a set of fit parameters obeys

$$P(\pi, \tau, \lambda, \epsilon \,|\{n_C\}, \{n_T\}) = \tag{14}$$
$$P(\{n_C\}, \{n_T\} \,|\, \pi, \tau, \lambda, \epsilon) \frac{P(\pi, \tau, \lambda, \epsilon)}{P(\{n_C\}, \{n_T\})}.$$

Here $P(\pi, \tau, \lambda, \epsilon)$ is the prior probability distribution, which we take to be uniform. In that case

$$P(\pi, \tau, \lambda, \epsilon | \{n_C\}, \{n_T\}) \tag{15}$$
$$\propto P(\{n_C\}, \{n_T\} \,|\, \pi, \tau, \lambda, \epsilon).$$

The parameter values with maximal likelihood are therefore those that maximize $P(\{n_C\}, \{n_T\} \,|\, \pi, \tau, \lambda, \epsilon)$.

In practice, it is more convenient to work with the logarithm of the likelihood, as

$$\log\left(P(\{n_C\}, \{n_T\} | \pi, \tau, \lambda, \epsilon)\right) \tag{16}$$
$$= \log\left(\prod_{n \in \{n_C\}} P_C(n) \prod_{n' \in \{n_T\}} P_T(n')\right)$$
$$= \sum_{n \in \{n_C\}} \log\left(P_C(n)\right) + \sum_{n' \in \{n_T\}} \log\left(P_T(n')\right).$$

14

If we define

$$X_C^{\leq} \equiv \sum_{n \in \{n_C\}, n \leq 2\tau} n,$$

$$X_C^{>} \equiv \sum_{n \in \{n_C\}, n > 2\tau} n,$$

$$X_T^{\leq} \equiv \sum_{n \in \{n_T\}, n \leq \tau + \pi} n,$$

$$X_T^{>} \equiv \sum_{n \in \{n_T\}, n > \tau + \pi} n.$$

and call the total number of convergent and tandem pairs N_C and N_T, this reduces to

$$\begin{aligned}
\log(P(\{n_C\}, \{n_T\} | \pi, \tau, \lambda, \epsilon)) & \quad (17)\\
= N_C \log(c_1) + N_T \log(c_2) & \\
- (\lambda - \epsilon)(X_C^{\leq} + X_T^{\leq}) & \\
- \lambda(X_C^{>} + X_T^{>}) & \\
+ 2N_C \epsilon \tau + N_T \epsilon (\tau + \pi), &
\end{aligned}$$

which can be maximized straightforwardly. To avoid possible influences of rare outliers, we only used values of n that fall in the domain that is plotted Fig. 5. This is correct if we modify c_1 and c_2 in equation 17 such that, given the domain D,

$$\int_D P_C(n)dn = \int_D P_T(n)dn = 1. \quad (18)$$

If we plot the likelihood as a function of one of the parameters, while keeping the other parameters at their maximum likelihood value, the plot can very well be approximated by a Gaussian. We use the standard deviation of this Gaussian as the error in the maximum likelihood parameter values.

In the main text, we discussed the values of τ and π, but not of q. The probability that two randomly chosen base pairs are the same and could therefore overlap is $\frac{1}{4}$. The fact that q is much higher than $\frac{1}{4}$ shows that "overlap" is much easier than expected based on this argument. This could reflect the density of functional elements, but also the flexibility of functional sequences, and the fact that regulatory regions are not mono-disperse.

D   Statistical tests

In this section we describe the statistical tests that are mentioned in the main text.
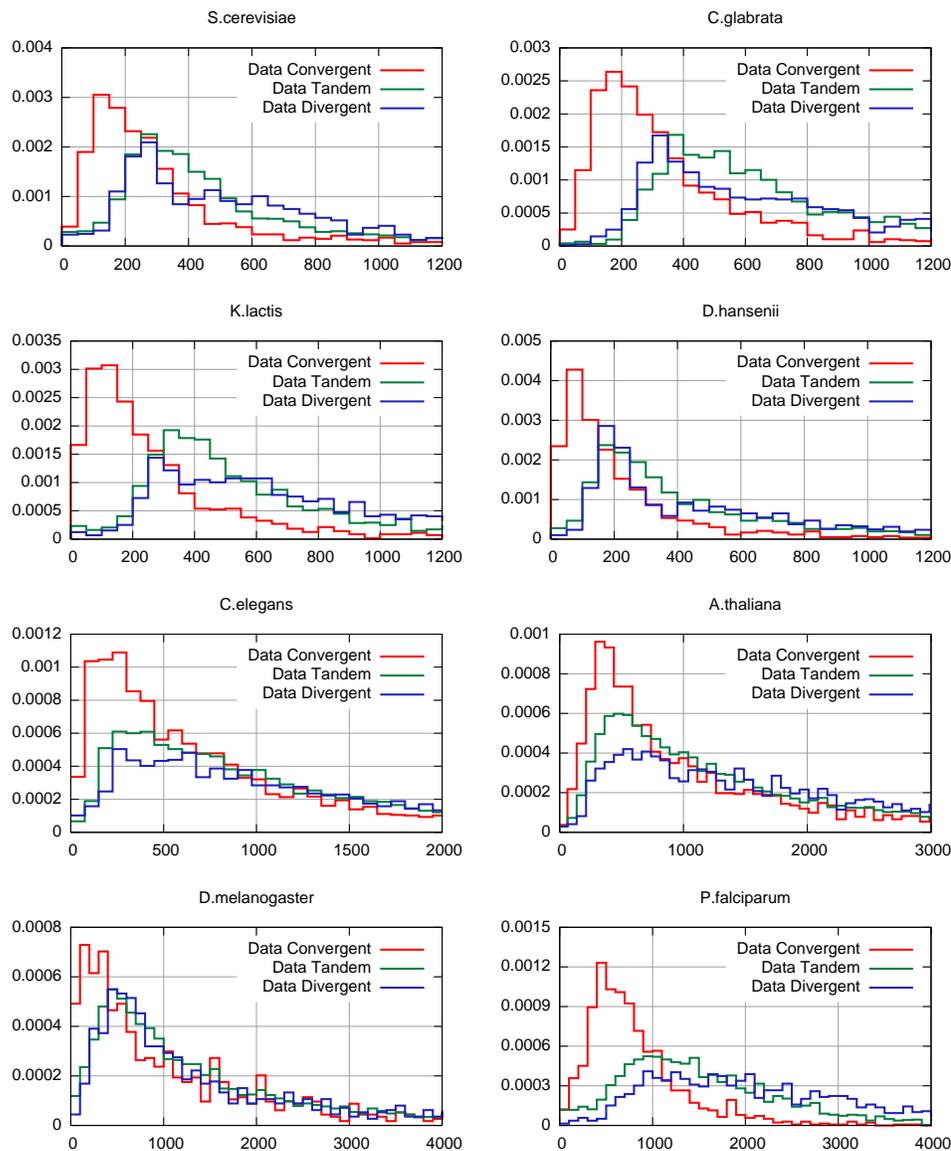
Figure 4: Distributions of intergenic distances, broken down into three different subsets (convergent, tandem, divergent pairs), for four fungi and four additional eukaryotes. Consistently, the convergent gene pairs are, on average, closer together than the tandem ones. The divergent genes are furthest apart. Note also that the divergent distribution is bimodal for all fungi, suggesting the presence of bi-directional promoters in all of them.
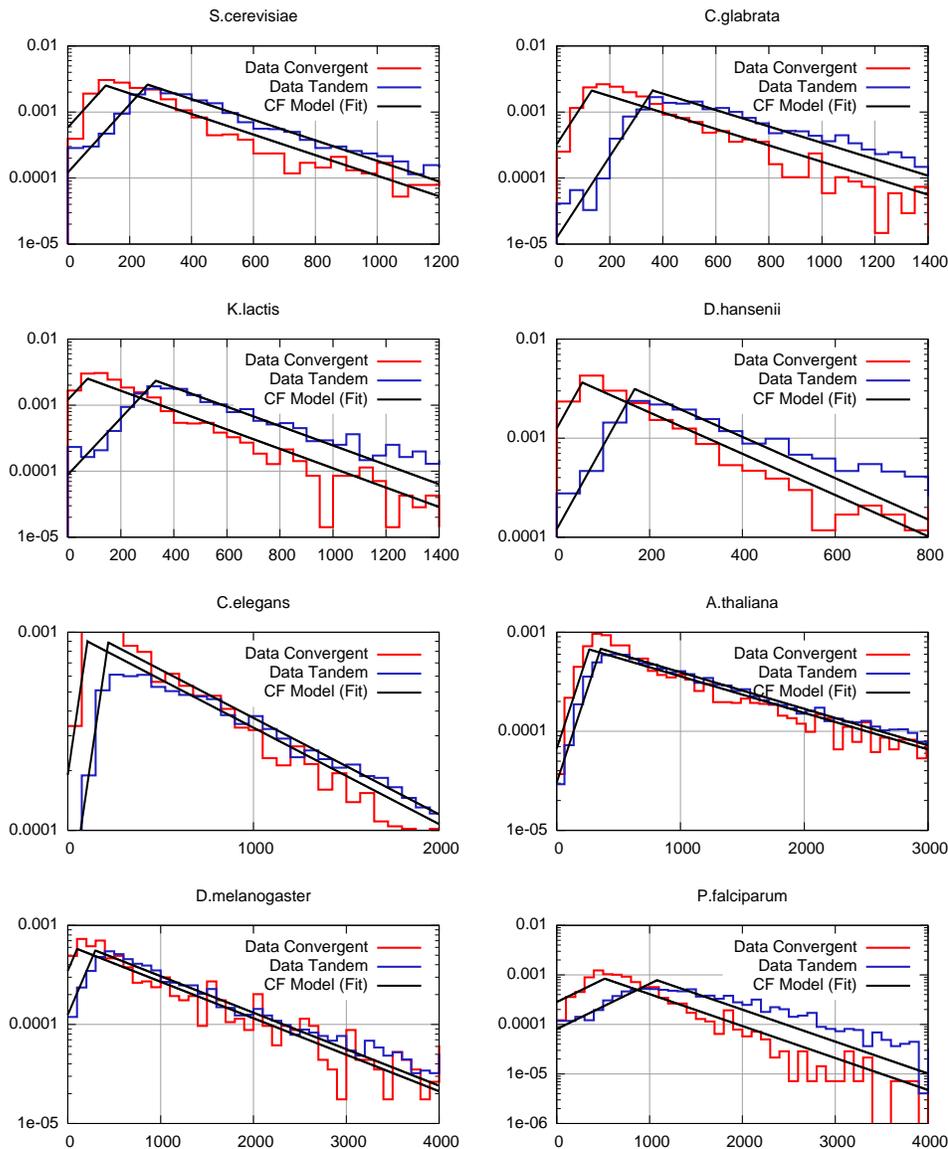
16

Figure 5: Maximum likelihood fits of the CF model to the distance distributions of four fungi and four additional eukaryotes. Despite the simplicity of the CF model, it does capture the qualitative features of each of the genomes, such as the dependence of the ORF spacing on relative orientation, and the exponential tails. The fits are used to estimate the sizes of UCRs and DCRs for the fungi; see Table 1 in the main text. Such quantitative estimates are probably not reliable for the higher eukaryotes, for which the model assumptions may be too crude.
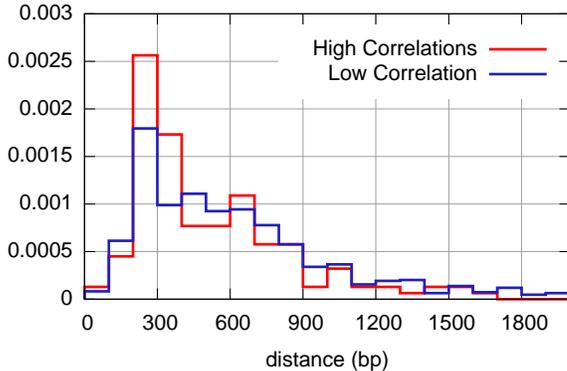
Figure 6: Distance distributions for divergent, neighboring gene pairs with high or low correlation in their expression, in *S. cerevisiae*. The figure shows that pairs with a high correlations coefficient are more likely to be close together than the ones with low correlation coefficients. This fact supports the hypothesis that the pairs in the first peak have a shared, bidirectional *cis*-regulatory region.

D.1   Co-expressed divergent pairs in *S. cerevisiae* are closer together than expected

In the main text, we state that co-expressed divergent gene pairs in *S. cerevisiae* have a tendency to have short intergenic regions. We tested this hypothesis as follows.

We used the expression data compiled by Dr. Andre Boorsma and Prof. Harmen J. Bussemaker to compute, for each neighboring pair of genes, the Pearson's correlation coefficient of their expression in about 900 experiments (see references [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38] for the original publications). These coefficients were usually low (< 0.3). We then split the set of divergent pairs into two subsets: those with a low correlation coefficient (< 0.3), called set 1, and those with a high one (≥ 0.3), called set 2. Next, we used a rank sum test to check whether the intergenic regions in set 2 are indeed shorter than expected at random.

The rank sum test was performed as follows. We first ranked the intergenic regions according to their length. Then, we computed the sum of the ranks of the intergenic regions in set 2; we call it R. Next, we randomized the ranks in the data set 10^7 times, each time computing the rank sum of set 2 in the randomized data. Finally, we counted the number of times R was smaller or equal to the rank sums obtained from the randomized data. The results show that the ORF pairs with a high correlation coefficient are significantly closer together than the ones with a low one ($p < 0.002$).

We checked that the observed signal is not due to paralogous gene pairs by excluding them from the set and repeating the test; this did not change the result. As a control experiment, we tested whether the same signal is also present in the set of *tandem* neighbors.

This is not the case (rank sum test: $p = 0.56$). We note, however, that a similar signal *was* present in the set of convergent pairs ($p < 2 \times 10^{-4}$); we have no satisfactory explanation for this fact.

The mean intergenic distance in set 1 is 721 bp; in set 2, it is 558 bp. The difference in the distance distributions of both sets is visually apparent (see Fig. 6).

D.2 Divergent genes in *S. cerevisiae* are more likely to be associated with the same process or component if they are close together

We studied whether there is an association between intergenic distance and functional similarity in divergent gene pairs in *S. cerevisiae*. In order to do this, we need to be able to quantify the functional similarity between two given genes. For this purpose we used the GO annotations of the GO Consortium (version 5.463 of Aug. 22 2007, see reference [12]) and the information-theoretic measure for semantic similarity proposed by Resnik [13].

The Gene Ontology is a hierarchical vocabulary of terms that can be assigned to genes. It falls apart into three independent taxonomies, each defined to describe one aspect of genes: the *biological process* they are involved in, the *molecular function* they perform and the *cellular component* they are active in.

The semantic similarity measure of Resnik provides, for each pair of GO terms a and b, a similarity score $s(a, b)$. If, for a given aspect α (biological process, molecular function or cellular component), gene A has been assigned GO terms $a_1 \ldots a_n$, and gene B has been assigned term $b_1 \ldots b_m$, then we define the similarity between genes A and B on this aspect as:

$$S(A, B, \alpha) = \max_{i,j} s(a_i, b_j). \tag{19}$$

Thus we can compute similarity values for each gene pair and for each aspect of the GO ontology.

If a certain gene did not have any assignment for some aspect, then the similarity score with any other gene was considered undefined on this aspect, and the gene was excluded from the analysis corresponding to this aspect.

We performed the following statistical tests. First, we divided the data set in two subsets, set 1 and set 2. The first set consisted of all divergent pairs that were in the "first" peak and set 2 contained all divergent neighbors that were further apart. As the bordering value we chose $d = 2\pi = 375$, since intergenic regions that are longer than that value can easily accommodate two independent promoters. We applied a Wilcoxon-Mann-Whitney rank-sum test to challenge the null hypothesis that the similarity scores of the pairs in set 1 and set 2 are drawn from the same distribution. We repeated this test for each aspect of the GO. The test results are $p = 8.8 \times 10^{-4}$, $p = 0.06$ and $p = 4.8 \times 10^{-5}$ for biological process, molecular function and cellular component respectively. We also ran a Spearman rank correlation test on the data, which resulted in the values $p = 2.0 \times 10^{-5}$, $p = 0.24$ and $p = 1.3 \times 10^{-5}$.

We conclude that the similarity scores belonging to the aspects biological process and cellular component are associated with intergenic distance. We do not, however, find a significant association between molecular function and intergenic distance. This is not very surprising, as proteins with a similar molecular function (e.g. "DNA binding" proteins), can act in very different processes and cellular components, so that there is no clear a priori reason to co-regulate them using a shared UCR.

We repeated this analysis for the convergent and tandem gene pairs. For the convergent pairs, none of the statistics were significant. However, the tandem pairs showed a similar pattern as the divergent ones; the Spearman rank correlation test resulted in $p = 4.6 \times 10^{-5}$, $p = 0.41$ and $p = 1.8 \times 10^{-5}$ for biological process, molecular function and cellular component respectively.

At this point, it is illustrative to point at one interesting example in Yeast: the BIO3, BIO4 and BIO5 cluster [39]. All genes in this cluster are involved in the biotin biosynthesis pathway. BIO3 and BIO4 are transcribed in a divergent orientation from a short intergenic region of length 222 bp (which falls into the first peak in the length distribution of divergent intergenic regions) and are tightly co-expressed. The orthologs of BIO3 and BIO4 in *E. coli* are BioA and BioB; these genes are closely spaced divergent neighbors as well (87 bp), and are simultaneously repressed by BirA binding to their shared UCR. Phalip *et al.* already speculate that a similar mechanism is at work in *S. cerevisiae* [39], but the mechanism of co-expression of these genes has not been studied in detail. BIO4 and BIO5 are tandem neighbors, and are only 55 bp apart, Clearly, this cluster has many of the features that we see in our statistical analysis. We therefore suggest that detailed experimental work on the regulation of the BIO cluster might illuminate some important mechanisms that shape the distribution of genes over the *S. cerevisiae* chromosome.

D.3 Greater than expected number of convergent intergenic regions with length 20-60 bp in *E. coli*

Here we show that the number of convergent intergenic regions with a length in the range 20-60 bp, is significantly larger than expected in *E. coli*.

In the calculation below, we estimate the significance of the peak in a very conservative way. We take the best-fitting exponential probability distribution as our null distribution (the fit is shown in Fig. 1b in the main text; its scale factor equals 145 bp). This way, we *underestimate* the statistical significance of the peak as we ignore the fact that the CF model actually predicts a *dip* in the distribution at the place of the peak.

Given the exponential null distribution, the fraction of the sample that is expected in the domain 20-60 bp is 0.21. Since the total number of convergent pairs is 543, the number of pairs in this domain is a random variable X that is distributed binomially with $p = 0.21$ and $N = 543$. The observed number of pairs in this domain in *E. coli* is 198; the probability for this to happen given the null distribution is $P(x \geq 198; p = 0.21, n = 543) < 10^{-13}$.

Based on the numbers above we should have expected $0.21 \times 543 = 114$ pairs in the

domain 20-60 bp. The actual observed number is 198; this means that we need about 86 bi-directional terminators to explain the data. If we assume that *E. coli* has 750 operons, we estimate that at least $\frac{2 \times 86}{750} \times 100\% = 23\%$ of the operons is terminated by a bi-directional terminator.

D.4 In *E. coli*, putative terminators in C regions are more often bi-directional than those in T regions

Here we show that the fraction of putative terminators that is classified as bi-directional by RNAMotif software [2], is larger in C regions than in T regions.

The statistical test was performed as follows. Our null hypothesis is that the terminators in the C region are a random sample from the total set of terminators in C or T regions. In total, the C and T regions together contain 1198 putative terminators, of which 222 are classified as bi-directional by RNAMotif. The C regions contain 378 putative terminators, of which 104 are bi-directional according to RNAMotif. If the 378 are chosen at random from the total set of 1198 terminators, then the number of terminators in the sample that are classified as bi-directional is a hypergeometric random variable. The probability to observe at least 104 bi-directional terminators in a random sample of 378 terminators, taken from a set of 1198 terminators containing 222 bi-directional ones, equals 1×10^{-7}.

D.5 Putative bi-directional terminators in C regions tend to occur in short regions

The fraction of the putative terminators in convergent intergenic regions that could be bi-directional (according to the RNAMotif algorithm) is significantly larger in short intergenic regions (< 100bp) than in long ones. We used the same statistical test as in the previous subsection. In total, the C regions contain 378 putative terminators; of these, 104 are classified as bi-directional. The short convergent intergenic regions contain 158 putative terminators, of which 62 are bi-directional according to RNAMotif. The probability to observe at least 62 bi-directional terminators in a random sample of 158 terminators, taken from a set of 378 terminators containing 104 bi-directional ones, is 2×10^{-5}.

D.6 Convergent operons that are close together are not more often functionally related

We tested whether the convergent operons that are close together (< 70bp) are more likely to be active in the same biological process or cellular component than ones that are further apart. For this we used the GO annotations from the GOA Database [40](version date: September 9. 2007). The same method was used as in section D.2, except that we now had to perform the analysis on the level of operons rather than genes. In order to compute the similarity between two operons, we compared the GO assignments for each gene in the first operon with each gene in second; the maximum of these scores was used as a similarity

measure for the operons. We did not find a significant signal for any of the aspects of the Gene Ontology (molecular function: $p = 0.20$; biological process: $p = 0.25$; cellular component: $p = 0.27$).

References

[1] P. B. Warren and P. R. ten Wolde, "Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*," *J. Mol. Biol.*, vol. 342, pp. 1379–1390, 2004.

[2] E. A. Lesnik, R. Sampath, H. B. Levene, T. J. Henderson, J. A. McNeil, and D. J. Ecker, "Prediction of rho-independent transcriptional terminators in escherichia coli," *Nucleic Acids Research*, vol. 29, no. 17, pp. 3583–3594, 2001.

[3] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome.," *Nature*, vol. 431, no. 7004, pp. 99–104, 2004.

[4] J. van Helden, M. del Olmo, and J. E. Perez-Ortin, "Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals.," *Nucleic Acids Res*, vol. 28, no. 4, pp. 1000–1010, 2000.

[5] J. H. Graber, G. D. McAllister, and T. F. Smith, "Probabilistic prediction of saccharomyces cerevisiae mrna 3'-processing sites.," *Nucleic Acids Res*, vol. 30, no. 8, pp. 1851–1858, 2002.

[6] L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz, "A high-resolution map of transcription in the yeast genome," *Proceedings of the National Academy of Sciences*, vol. 103, no. 14, pp. 5320–5325, 2006.

[7] I. Tirosh, J. Berman, and N. Barkai, "The pattern and evolution of yeast promoter bendability," *Trends in Genetics*, vol. 23, no. 7, pp. 318–321, 2007.

[8] P. J. Bell, P. H. Bissinger, R. J. Evans, and I. W. Dawes, "A two-reporter gene system for the analysis of bi-directional transcription from the divergent mal6t-mal6s promoter in saccharomyces cerevisiae.," *Curr Genet*, vol. 28, no. 5, pp. 441–446, 1995.

[9] Y. Liu and W. Xiao, "Bidirectional regulation of two dna-damage-inducible genes, mag1 and ddi1, from saccharomyces cerevisiae.," *Mol Microbiol*, vol. 23, no. 4, pp. 777–789, 1997.

[10] C. Aranda, M. Colon, C. Ishida, L. Riego, A. Deluna, L. Valenzuela, J. Herrera, and A. Gonzalez, "Gcn5p contributes to the bidirectional character of the uga3-glt1 yeast promoter.," *Biochem Biophys Res Commun*, vol. 348, no. 3, pp. 989–996, 2006.

[11] C. Ishida, C. Aranda, L. Valenzuela, L. Riego, A. Deluna, F. Recillas-Targa, P. Filetici, R. Lopez-Revilla, and A. Gonzalez, "The uga3-glt1 intergenic region constitutes a promoter whose bidirectional nature is determined by chromatin organization in saccharomyces cerevisiae.," *Mol Microbiol*, vol. 59, no. 6, pp. 1790–1806, 2006.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nat Genet*, vol. 25, no. 1, pp. 25–29, 2000.

[13] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, pp. 95–130, 1998.

[14] K. Postle and R. F. Good, "A bidirectional rho-independent transcription terminator between the e. coli tonb gene and an opposing gene;," *Cell*, vol. 41, no. 2, pp. 577–585, 1985.

[15] N. Yachie, K. Arakawa, and M. Tomita, "On the interplay of gene positioning and the role of rho-independent terminators in escherichia coli.," *FEBS Lett*, vol. 580, no. 30, pp. 6909–6914, 2006.

[16] H. Salgado, G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides, "Operons in escherichia coli: Genomic analyses and predictions," *Proc. Natl. Acad. Sci. USA*, vol. 97, no. 12, pp. 6652–6657, 2000.

[17] L. Tonks, "The complete equation of state of one, two and three-dimensional gases of hard elastic spheres," *Physical Review*, vol. 50, no. 10, 1936.

[18] F. Perocchi, Z. Xu, S. Clauder-Munster, and L. M. Steinmetz, "Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin d," *Nucleic Acids Research*, pp. gkm683–, 2007.

[19] A. Hajarnavis, I. Korf, and R. Durbin, "A probabilistic model of 3' end formation in caenorhabditis elegans.," *Nucleic Acids Res*, vol. 32, no. 11, pp. 3392–3399, 2004.

[20] V. M. Boer, J. H. de Winde, J. T. Pronk, and M. D. W. Piper, "The genome-wide transcriptional responses of saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur.," *J Biol Chem*, vol. 278, no. 5, pp. 3265–3274, 2003.

[21] A. Boorsma, H. de Nobel, B. ter Riet, B. Bargmann, S. Brul, K. J. Hellingwerf, and F. M. Klis, "Characterization of the transcriptional response to cell wall stress in saccharomyces cerevisiae.," *Yeast*, vol. 21, no. 5, pp. 413–427, 2004.

[22] C. Bro, B. Regenberg, G. Lagniel, J. Labarre, M. Montero-Lomeli, and J. Nielsen, "Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells.," *J Biol Chem*, vol. 278, no. 34, pp. 32141–32149, 2003.

[23] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast.," *Science*, vol. 282, no. 5389, pp. 699–705, 1998.

[24] P. Daran-Lapujade, M. L. A. Jansen, J.-M. Daran, W. van Gulik, J. H. de Winde, and J. T. Pronk, "Role of transcriptional regulation in controlling fluxes in central carbon metabolism of saccharomyces cerevisiae. a chemostat culture study.," *J Biol Chem*, vol. 279, no. 10, pp. 9125–9138, 2004.

[25] F. Devaux, P. Marc, C. Bouchoux, T. Delaveau, I. Hikkel, M. C. Potier, and C. Jacq, "An artificial transcription activator mimics the genome-wide properties of the yeast pdr1 transcription factor.," *EMBO Rep*, vol. 2, no. 6, pp. 493–498, 2001.

[26] J. A. Fleming, E. S. Lightcap, S. Sadis, V. Thoroddsen, C. E. Bulawa, and R. K. Blackman, "Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by ps-341.," *Proc Natl Acad Sci U S A*, vol. 99, no. 3, pp. 1461–1466, 2002.

[27] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown, "Genomic expression programs in the response of yeast cells to environmental changes.," *Mol Biol Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[28] A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, and P. O. Brown, "Genomic expression responses to dna-damaging agents and the regulatory role of the yeast atr homolog mec1p," *Mol Biol Cell*, vol. 12, no. 10, pp. 2987–3003, 2001.

[29] K. Harris, R. E. Lamson, B. Nelson, T. R. Hughes, M. J. Marton, C. J. Roberts, C. Boone, and P. M. Pryciak, "Role of scaffolds in map kinase pathway specificity revealed by custom design of pathway-dedicated signaling proteins.," *Curr Biol*, vol. 11, no. 23, pp. 1815–1824, 2001.

[30] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional discovery via a compendium of expression profiles.," *Cell*, vol. 102, no. 1, pp. 109–126, 2000.

[31] A. Lagorce, N. C. Hauser, D. Labourdette, C. Rodriguez, H. Martin-Yken, J. Arroyo, J. D. Hoheisel, and J. Francois, "Genome-wide analysis of the response to cell wall mutations in the yeast saccharomyces cerevisiae.," *J Biol Chem*, vol. 278, no. 22, pp. 20345–20357, 2003.

[32] M. T. McCammon, C. B. Epstein, B. Przybyla-Zawislak, L. McAlister-Henn, and R. A. Butow, "Global transcription analysis of krebs tricarboxylic acid cycle mutants reveals an alternating pattern of gene expression and effects on hypoxic and oxidative genes.," *Mol Biol Cell*, vol. 14, no. 3, pp. 958–972, 2003.

[33] S. Mnaimneh, A. P. Davierwala, J. Haynes, J. Moffat, W.-T. Peng, W. Zhang, X. Yang, J. Pootoolal, G. Chua, A. Lopez, M. Trochesset, D. Morse, N. J. Krogan, S. L. Hiley, Z. Li, Q. Morris, J. Grigull, N. Mitsakakis, C. J. Roberts, J. F. Greenblatt, C. Boone, C. A. Kaiser, B. J. Andrews, and T. R. Hughes, "Exploration of essential gene functions via titratable promoter alleles.," *Cell*, vol. 118, no. 1, pp. 31–44, 2004.

[34] Y. Murata, T. Watanabe, M. Sato, Y. Momose, T. Nakahara, S.-i. Oka, and H. Iwahashi, "Dimethyl sulfoxide exposure facilitates phospholipid biosynthesis and cellular membrane proliferation in yeast cells.," *J Biol Chem*, vol. 278, no. 35, pp. 33185–33193, 2003.

[35] T. Sahara, T. Goda, and S. Ohgiya, "Comprehensive expression analysis of time-dependent genetic responses in yeast cells to low temperature.," *J Biol Chem*, vol. 277, no. 51, pp. 50015–50021, 2002.

[36] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization.," *Mol Biol Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[37] S. L. Tai, V. M. Boer, P. Daran-Lapujade, M. C. Walsh, J. H. de Winde, J.-M. Daran, and J. T. Pronk, "Two-dimensional transcriptome analysis in chemostat cultures. combinatorial effects of oxygen availability and macronutrient limitation in saccharomyces cerevisiae.," *J Biol Chem*, vol. 280, no. 1, pp. 437–447, 2005.

[38] H. Yoshimoto, K. Saltsman, A. P. Gasch, H. X. Li, N. Ogawa, D. Botstein, P. O. Brown, and M. S. Cyert, "Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in saccharomyces cerevisiae.," *J Biol Chem*, vol. 277, no. 34, pp. 31079–31088, 2002.

[39] V. Phalip, I. Kuhn, Y. Lemoine, and J. M. Jeltsch, "Characterization of the biotin biosynthesis pathway in saccharomyces cerevisiae and evidence for a cluster containing bio5, a novel gene involved in vitamer uptake.," *Gene*, vol. 232, no. 1, pp. 43–51, 1999.

[40] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology.," *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D262–6, 2004.