Alignments for the intergenic regions

For the cross-species comparisons we first define orthologous genes between the species in question. This is done by comparing NCBI gene tables. Then we extract promoter regions for the orthologous gene pairs by including 700 base pairs upstream (downstream) from the starting positions of the positive (negative) strand genes. If another gene is located closer than 700 base pairs we stop there. We then use CLUSTALW1.83 [1] with default parameters for the alignment. We also take into account inversions and test for each intergenic region whether such an event has taken place. These alignment libraries can be used to search for the binding sites of any sequence specific transcription factor since they are *independent* of the energy matrices. Following [2], we disregard a site pair if the alignment places gaps between them.

Background distribution P_0

For the genomic background distribution $P_0(\mathbf{a})$, we use the following conditional frequencies extracted from the statistics of nucleotide pairs (a', a) in *E. coli* intergenic regions and the corresponding single-nucleotide frequencies

$$\pi_0(a|a') = \begin{pmatrix} 0.3410 & 0.2230 & 0.1977 & 0.2381 \\ 0.2438 & 0.2227 & 0.2793 & 0.2540 \\ 0.2408 & 0.2354 & 0.2285 & 0.2951 \\ 0.2920 & 0.1823 & 0.1877 & 0.3378 \end{pmatrix}, \qquad p_0(a) = \begin{pmatrix} 0.2847 \\ 0.2141 \\ 0.2192 \\ 0.2818 \end{pmatrix}. \tag{S1}$$

The underlying mutation model is a special case of the class of models discussed in [3]. We create 10^8 sequences $(a_1, \ldots, a_\ell)$ according to this model, which are then used to build the distribution $P_0(E)$ with energy $E = \sum_{i=1}^{l} \epsilon_i(a_i)$. The neighbor dependence of the nucleotide frequencies proves important to describe the low-energy tail of $P_0(E)$ correctly [4]. The energy matrix $\epsilon_i(a)$ is obtained using the position weight matrix method [5]. Standard pseudocounts of $+1$ are used in the energy matrix construction to avoid singularities due to finite sample size (we use 48 experimentally verified sites [6]). We use this energy matrix for the projection of the distribution $P_0(\mathbf{a})$ onto the energy variable E. All following steps of the analysis need to be performed for the given transcription factor under study, in contrast to the alignment libraries and the genomic background distribution which are independent of the energy matrices. It is also important to check that the transcription factor in question is conserved between the species compared in order to justify the usage of a single energy matrix.

Energy transition probabilities G_0^t and G_s^t

The energy transition probability distribution for neutral evolution G_0^t is constructed by generating a set of 1000 independent sequence states for each energy level E_i. Energy levels run from $E_{\min} = 0.0$ to

$E_{\max} = 35.0$ with intervals $dE = 0.1$. We then evolve these states over an evolutionary distance t to obtain the distribution $G_0^t(E_2|E_1)$. For this purpose, it proves sufficient to approximate the neighbor dependent substitution model by a single-nucleotide model of the Kimura two-parameter form [7]. The Kimura transition matrix forms the link between the observed sequence similarities and the evolutionary distances. The ratio between transitions and transversions is estimated from the statistics of the alignment of the intergenic regions to be about 1.5. The evolutionary distance t is extracted from the alignment of the orthologous site pairs [7]. In a most parsimonious approximation, we use a single value of t for all intergenic regions. In order to evaluate expressions which include integrations over the time variable, such as eq. (13), we need to compute $G_0^t(E_2|E_1)$ for a set of evolutionary distances t.

The transition probabilities $G_s^t(E_2|E_1)$ are simulated in the same way; the substitution rates are modified with respect to the neutral case as given by eq. (1). These rates are now position-dependent. Moreover, as stated in the main text, the rate of fixation of a mutation a in a position i depends on all the other $l-1$ positions. The fitness function is extracted, as explained in the main text, by first forming the distributions $Q(E)$ and $P_0(E)$ and then applying eq. (6), see fig. 3(b). We also apply Gaussian kernel estimators for the G functions to reduce noise.

Three-species comparisons

In the alignments of the three species compared, we have 52527 orthologous sites pairs between *E. coli* and *Y. pseudotuberculosis*, 54416 between *S. typhimurium* and *Y. pseudotuberculosis*, and finally 10128 triplets between all three genomes. The construction of the triplets via pairwise alignments produces a conservative set of triplets biased towards conservation, as given by the parameter $\lambda = 0.0035$. Plotting the energy triplets (E_1, E_2, E_3) of aligned loci yields a distribution with a well-conserved low-energy tail as expected (see fig. 6, which is published as Supporting Information on the PNAS website). Improving the alignment will produce more candidate loci with functional innovations. This is an important issue but is beyond the scope of this paper.

The background phylogeny of the three species is obtained using a distance-based method [7]. We find the ratios 1 : 2.2 : 1.9 for the pairwise distances *E. coli* - *S. typhimurium*, *E. coli* - *Y. pseudotuberculosis*, and *S. typhimurium* - *Y. pseudotuberculosis*. Assuming the root to be the midpoint of the tree ($t_3 = (t_1 + t_2)/2$), this defines the distances of the three species from the root, $\mathbf{t} \equiv (t_1, t_2, t_3)$, up to an overall scale. We can now build functional phylogenies generalizing the definitions in the main text. Considering again only cases with at most one functional switch, the hidden Markov model for the three-species energy data is of the form

$$W^{\mathbf{t}}(E_1, E_2, E_3) = \sum_{\alpha_1 \in 0,s} \sum_{\alpha_2 \in 0,s} \sum_{\alpha_3 \in 0,s} \lambda^{\mathbf{t}}_{\alpha_1,\alpha_2,\alpha_3} R^{\mathbf{t}}_{\alpha_1,\alpha_2,\alpha_3}(E_1, E_2, E_3). \tag{S2}$$

The eight conditional distributions are $R^{\mathbf{t}}_{000} \equiv P^{\mathbf{t}}_0$ (neutral evolution), $R^{\mathbf{t}}_{00s}$, $R^{\mathbf{t}}_{0s0}$, and $R^{\mathbf{t}}_{s00}$ (time-dependent selection with a functional locus in one species), $R^{\mathbf{t}}_{0ss}$, $R^{\mathbf{t}}_{s0s}$, and $R^{\mathbf{t}}_{ss0}$ (time-dependent selection with a functional locus in two species), and $R^{\mathbf{t}}_{sss} \equiv Q^{\mathbf{t}}$ (time-independent selection, conserved functionality). We obtain for neutral evolution

$$\begin{aligned} P^{\mathbf{t}}_0(E_1, E_2, E_3) &= \int dE_a dE_f G_0^{t_3}(E_3|E_a) G_0^{t_f}(E_f|E_a) G_0^{t_1-t_f}(E_1|E_f) G_0^{t_2-t_f}(E_2|E_f) P_0(E_a) \qquad \text{(S3)} \\ &= \int dE_f G_0^{t_3+t_f}(E_3|E_f) G_0^{t_1-t_f}(E_1|E_f) G_0^{t_2-t_f}(E_2|E_f) P_0(E_f). \end{aligned}$$

Each branch is represented by a transition probabillity G_0, and the unobserved energies E_a at the root and E_f at the internal node of the tree are integrated over. The second equality follows from detailed balance, which enables us to eliminate the integration over the ancestral energy variable E_a. The distribution $Q(E_1, E_2, E_3)$ for conserved selection takes a similar form with transition probabilities G_s instead of G_0 and the distribution Q instead of P_0.

For the ensembles with time-dependent selection, we obtain, e.g.,

$$
\begin{aligned}
R^{\mathbf{t}}_{00s}(E_1,E_2,E_3) &= \frac{1}{\lambda^{\mathbf{t}}_{00s}} \int dE' dE_a dE_f \times \qquad \text{(S4)}\\
\Big[(1-\lambda)\nu_+ &\int_0^{t_3} dt' G_s^{t_3-t'}(E_3|E')G_0^{t'}(E'|E_a)G_0^{t_f}(E_f|E_a)G_0^{t_1-t_f}(E_1|E_f)G_0^{t_2-t_f}(E_2|E_f)P_0(E_a)\\
+\lambda\nu_- &\int_0^{t_f} dt' G_s^{t_3}(E_3|E_a)G_s^{t'}(E'|E_a)G_0^{t_f-t'}(E_f|E')G_0^{t_1-t_f}(E_1|E_f)G_0^{t_2-t_f}(E_2|E_f)Q(E_a)\Big]\\
&= \frac{1}{\lambda^{\mathbf{t}}_{00s}} \int dE' dE_f \times\\
\Big[(1-\lambda)\nu_+ &\int_0^{t_3} dt' G_s^{t_3-t'}(E_3|E')G_0^{t'+t_f}(E'|E_f)G_0^{t_1-t_f}(E_1|E_f)G_0^{t_2-t_f}(E_2|E_f)P_0(E_f)\\
+\lambda\nu_- &\int_0^{t_f} dt' G_s^{t_3+t'}(E_3|E')G_0^{t_f-t'}(E_f|E')G_0^{t_1-t_f}(E_1|E_f)G_0^{t_2-t_f}(E_2|E_f)Q(E')\Big],
\end{aligned}
$$

with $\lambda^{\mathbf{t}}_{00s} = (1-\lambda)\nu_+ t_3 + \lambda\nu_- t_f$ and $t_f = (t_1 + t_2 - \Delta_{12})/2$, where Δ_{12} is the distance between species 1 and 2 (see fig. 7, which is published as Supporting Information on the PNAS website). We have again eliminated the integration over E_a using detailed balance. The first term of eq. (S4) corresponds to an ancestral locus under neutral evolution, which then at some point t' gains functionality due to changed selection pressure. Similarly, the second term describes loss of functionality due to a change in selection pressure at time t'. The prefactor $\lambda^{\mathbf{t}}$ in eq. (S2) contains the prior weight of the ensemble R_{00s}, i.e, the combined probability of a functional (nonfunctional) ancestral site and a subsequent loss (gain) of function. The prefactors in eq. (S4) describe the individual weights of these two alternatives within the R_{00s} ensemble. With increasing number of species, it will potentially be useful to further differentiate between the gain and loss of functionality events. This is straightforward to do and results in displaying the separate terms of eq. (S4) as distinct ensembles in eq. (S2). For the three species case considered here, this is not a statistically attractive option.

Now we are in a position to proceed exactly as in the two species case and evaluate probabilities for each energy triplet analogously to eq. (15),

$$
\rho^{\mathbf{t}}_{\alpha_1,\alpha_2,\alpha_3}(E_1,E_2,E_3) = \frac{\lambda^{\mathbf{t}}_{\alpha_1,\alpha_2,\alpha_3} R^{\mathbf{t}}_{\alpha_1,\alpha_2,\alpha_3}(E_1,E_2,E_3)}{W^{\mathbf{t}}(E_1,E_2,E_3)}. \qquad \text{(S5)}
$$

The prediction lists are available upon request. Out of the 11 verified CRP binding sites in the orthologous energy triplets, we predict with high probability conserved functionality for 10 triplets. The "missing" verified site is the fourth *malE-malK* site discussed in the main text. These three species allow an almost clean disentanglement of the background and the functional loci (see fig. 5, which is published as Supporting Information on the PNAS website). However, as pointed out earlier, this comes at the price of a lower number of total predictions due to the conservative alignment procedure.

References

[1] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. (2003) *Nucleic Acids Res.* **31**, 3497-3500.

[2] Brown, C. T. & Callan, C.G. Jr, (2004) *Proc. Natl. Acad. Sci. USA* **101**, 2404-2409.

[3] Arndt, P. & Hwa, T. (2005) *Bioinformatics* **21**, 2322-2328.

[4] Djordjevic, M., Sengupta, A.M. & Shraiman, B. I. (2003) *Genome Res.* **13**, 2381-2390.

[5] Berg, O. & von Hippel, P. (1987) *J. Mol. Biol.* **193**, 723-750.

[6] Robison, K., McGuire, A. M. & Church, G. M. (1988) *J. Mol. Biol.* **284**, 241-254.

[7] Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. (1998) *Biological sequence analysis* (CUP, Cambridge, UK).