I. SEPARATION OF HIGH AND LOW CONSERVED REGIONS WITH A HIDDEN MARKOV MODEL

In general, an HMM is a model which assumes that there are hidden states (in our case, the two hidden states are neutral or under selective constraint) at each position in a sequence. The observed sequence values (in our case, these are the conservation patterns along the aligned promoter sequences) are probabilistic outcomes emitted by the hidden states. These emission probabilities, as well as the transition probabilities between hidden states, are unknown parameters which can be learned through an iterative procedure that attempts to maximize the likelihood of the observed sequence as a function of these parameters [Baum 1972]. The value of the hidden state at each position can be recovered by considering the sequence of hidden states most likely to have produced the observed sequence values.

We used a Markov model with two hidden states – high conservation region (HCR) and low conservation region (LCR) – to perform this separation. Gaps were implemented as explicit states in the Markov model but were ignored in the conservation rate analysis.

For each promoter, we used up to 600 base pairs of sequence upstream from the ATG start codon, avoiding any overlap with coding sequences. (Since our two-state HMM does not allow for multiple types of conserved regions, coding sequence will bias the inferred HMM parameters in ways unrelated to promoter conservation.) Promoter sequences shared by two ORFs were only used once. These criteria yielded 2453 promoter sequences. The promoters were aligned between species using CLUSTALW with default parameters.

The topology of the HMM is shown in Fig. 1. There were two hidden states, "L" and "H", and one explicit state "G". Along the *S. cerevisiae* promoters, we identified each site as one of three observed states: "match", "mismatch" and "gap". The "L" and "H" hidden states were allowed to emit "match" and "mismatch" as shown by the solid arrows, but the "G" state was only allowed to emit "gap". All transitions (indicated as dashed lines) between "L", "H", and "G" were allowed.

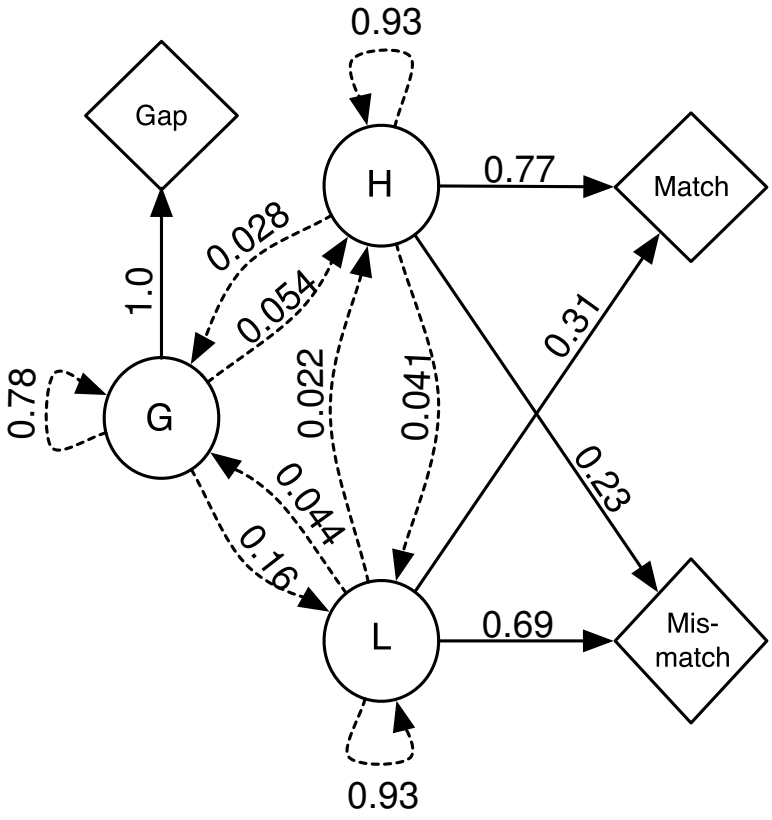

FIG. 1: **State diagram of the Hidden Markov Model used to separate high and low conservation regions.** There were two hidden states, "H" and "L", and one explicit state "G". Along the *S. cerevisiae* promoters, we identified each site as one of three observed states: "match", "mismatch" and "gap". The "H" and "L" hidden states were allowed to emit "match" and "mismatch" as shown by the solid arrows, but the "G" state was only allowed to emit "gap". The numerical values indicate the converged transition and emission probabilities inferred from the sequence data.

We trained the HMM using the standard Baum-Welch algorithm [Baum 1972, Durbin et al. 1998]. This is an iterative procedure that attempts to find the maximum likelihood values for the transition and emission parameters given the sequence of emitted states. The Viterbi algorithm was used to identify the most probable state path for the promoter sequences [Durbin et al. 1998]. This path was used to assign sites to HCR or LCRs. Note that if we did not know rates were uniform, a separate HMM would have been necessary for each promoter, which would have greatly increased the uncertainty in the results.

We applied the Baum-Welch algorithm to infer the transition and emission parameters using three different initial conditions. We iterated the likelihood maximization procedure until none of the parameters changed by more than 10^{-5} in a single iteration. For each initial condition, we inferred the same final parameters, all robust to within 10^{-5}, suggesting that local minima are not an issue in this dataset. The three initial conditions and the final parameters are shown below. The T matrix indicates the transition probabilities between hidden states, with the ordering of the states being (High, Low, Gap). The entry in row i and column j indicates the transition from state i to state j. The E matrix indicates the emission probabilities from a hidden state to an observable. There are three columns, corresponding to the hidden states High, Low, and Gap. The three rows correspond to the emitted observables Match, Mismatch, and Gap. The π vector indicates the probabilities that the hidden state that precedes the first site in each promoter alignment is High, Low, or Gap. The final estimated emission probabilities were $P(\text{match}|\text{HCR}) = 0.77$ and $P(\text{match}|\text{LCR}) = 0.31$. $P(\text{match}|\text{LCR})$ was very close to what we expected based on the 4-fold site statistics (0.33). The complete set of converged transmission and emission parameters are also shown in Fig. 1.

$$1)T_0 = \begin{pmatrix} 0.64 & 0.27 & 0.09 \\ 0.64 & 0.27 & 0.09 \\ 0.33 & 0.33 & 0.34 \end{pmatrix}, E_0 = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.7 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \pi_0 = \begin{pmatrix} 0.33 \\ 0.34 \\ 0.33 \end{pmatrix}$$

$$2)T_0 = \begin{pmatrix} 0.55 & 0.36 & 0.09 \\ 0.36 & 0.55 & 0.09 \\ 0.33 & 0.33 & 0.34 \end{pmatrix}, E_0 = \begin{pmatrix} 0.6 & 0.4 & 0 \\ 0.4 & 0.6 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \pi_0 = \begin{pmatrix} 0.33 \\ 0.34 \\ 0.33 \end{pmatrix}$$

$$3)T_0 = \begin{pmatrix} 0.55 & 0.36 & 0.09 \\ 0.36 & 0.55 & 0.09 \\ 0.33 & 0.33 & 0.34 \end{pmatrix}, E_0 = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \pi_0 = \begin{pmatrix} 0.33 \\ 0.34 \\ 0.33 \end{pmatrix}$$

$$T_f = \begin{pmatrix} 0.93 & 0.041 & 0.028 \\ 0.022 & 0.93 & 0.044 \\ 0.054 & 0.16 & 0.78 \end{pmatrix}, E_f = \begin{pmatrix} 0.77 & 0.31 & 0 \\ 0.23 & 0.69 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \pi_f = \begin{pmatrix} 0.05 \\ 0.05 \\ 0.90 \end{pmatrix}$$

The HMM would not have detected separate distributions if conservation rates were homogeneous in the promoters. To illustrate this, we trained the HMM on a conservation pattern generated from a uniform mutation model with conservation rate equal to the overall conservation rate in the promoters. In this synthetic case, the HMM failed to separate the promoters into two types of sequence. Almost all of the sequence was classified (98%) into one state.

We also tested whether stretches of conservation were artifacts of the alignment procedure by applying the alignment procedure and HMM to a set of artificial sequences. The artificial sequences were generated as follows. We counted the frequencies of all quadruplets $b_{scer}b_{spar}b_{smik}b_{sbay}$ occurring at shared 4-fold sites, where the b's indicate the nucleotide in each species. The quadruplet frequencies allowed us to determine the neutral probability $P(b_{spar}b_{smik}b_{sbay}|b_{scer})$ of observing bases $b_{spar}b_{smik}b_{sbay}$ given b_{scer}. We used $P(b_{spar}b_{smik}b_{sbay}|b_{scer})$ and the promoter sequences of S. cerevisiae to generate artificial promoter sequences for the three other species. This procedure allowed us to take account of phylogenetic information without having to explicitly choose an ancestral sequence or evolutionary model. We aligned these artificial sequences with ClustalW and applied the HMM to the alignments. The HMM found no separation of high and low conservation regions, showing that the long stretches of conservation in the real promoters were not alignment artifacts.

We found additional evidence suggesting the functional relevance of the HCRs by analyzing how conservation depends on location within promoters . For each value of x – the distance from the ATG – we measured a conservation fraction based on the number of promoters which had a conserved site at that position. Let $N_c(x)$ be the number of promoters that have a conserved site at the position x, and let $N_a(x)$ be the number of promoters that do not have a conserved site at position x. The conservation frequency at x is defined as $N_c(x)/(N_c(x) + N_a(x))$. Since each conserved site is either in an LCR or a HCR, we can decompose $N_c(x)$ as $N_c(x) = N_c^h(x) + N_c^l(x)$, where $N_c^h(x)$ and $N_c^l(x)$ are the numbers of promoters in which the site x is in an HCR or in an LCR, respectively. The contributions of HCR and LCR to the conservation frequency are defined as $N_c^h(x)/(N_c(x) + N_a(x))$ and $N_c^l(x)/(N_c(x) + N_a(x))$.

It is known that regulatory elements generally occur 100 to 400 bp upstream from the ATG start codon[Cliften et al. 2003]. Consistent with this fact, we observed that the fraction of promoters with conserved sites was higher in this region (black curve in Fig. 2). These higher conservation rates were due to contributions from the HCRs. In this region, there were more HCRs than elsewhere (inset in Fig. 2), and conserved sites were more often in HCRs (blue curve in main graph) than in LCRs (red curve).

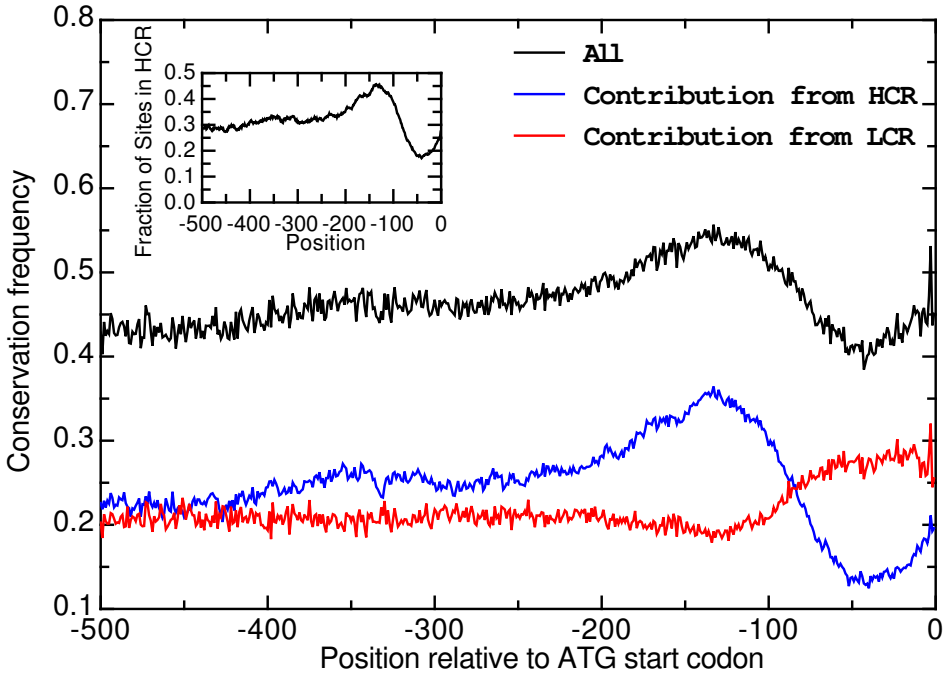

FIG. 2: **Figure 6. Conservation rate as a function of location.** For each position, the black curve shows the fraction of promoters where that site is conserved in all four species. The contributions to the conservation from bases in HCRs (blue curve) and LCRs (red curve) are also shown. The inset shows the fraction of sites that are in HCRs, as a function of location.

Although other regions (e.g. [-500 bp, -400 bp] and [-100 bp, 0]) had overall conservation rates comparable to each other, the proportion of HCRs and LCRs in these regions differed depending on the distance from the ATG. In the region nearest to the ATG ([-100,0]), LCR sites contributed more to the overall conservation than did the HCR sites, as illustrated by the crossover in the red and blue curves at -100 bp. This suggested that a more common behavior in the region [-100,0] is to have a moderate level of sparsely conserved sites, with fewer blocks that would be classified as an HCR. Meanwhile, for the region [-500,-400], it suggested that some promoters have blocks of conserved sites, while other promoters have few conserved sites. One explanation could be that the typical start of transcription relative to the translation start is around -100 bp. The near region may represent the 5' UTR, which could be subject to different selective pressures that would give rise to alternative patterns of conservation.

II. REFERENCES

[Baum 1972] Baum, L. E. 1972. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities.* **3**:1-8.

[Cliften et al. 2003] Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**:71-76.

[Durbin et al. 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis*, Cambridge University Press, Cambridge.