Appendix 1

Selection of model of amino acid substitution.

In the SOWHL tests performed on amino acid alignments either the WAG (Whelan and Goldman 2001) or the JTT (Jones et al. 1992) replacement matrix was used. The WAG matrix was selected over others on the basis that it was estimated from a large database of different globular protein families, as opposed being tailored to a specific family (as in the mtREV24 and mtmamm matrices), and it avoids some of the pitfalls of matrices derived from counting-based methods (such as the Dayhoff and JTT matrices, see Whelan and Goldman (Whelan and Goldman 2001) and references therein). The JTT matrix was used to test the assumptions of the SOWHL test (see main text). We used the second order Akaike Information Criterion (AIC_c) to select other aspects of the model, namely whether to estimate amino acid frequencies from the data and whether to include a gamma model of rate variation among sites (Hurvich and Tsai 1989). The Akaike Information Criterion (AIC) is a measure of the amount of information lost when a particular model is used to approximate a given situation. The second order AIC is used when the number of free parameters in the model is close to the number of sites in the alignment (Posada and Buckley 2004), as is the case here. The AIC_c is calculated as:

$$AIC_c = -2\ln L + 2K + \frac{2K(K+1)}{n-k-1}$$

where L is the likelihood of a given phylogenetic tree, K is the number of free parameters in the model being assumed and n is the number of sites in the alignment. An AIC_c score is calculated for each model under consideration. Models are then compared using the difference in AIC_c (Δ) calculated as follows:

$$\Delta_i = AIC_i - AIC_{\min}$$

where AIC_{min} is the lowest obtained AIC_c value among candidate models (i.e. the AIC_c of the best model). Models with a Δ_i >10 can be confidently rejected as a worse fit to the data than the best model (Posada and Crandall 1998).

Four different models of protein evolution were considered, using the alignment of homeodomains shown in figure 1 (main text). These were, WAG, WAG+F (which allows amino acid frequencies to be estimated from the data), WAG+Γ (which allows rates of evolution to vary across sites according to a gamma distribution with 8 categories), and the WAG+F+Γ (which allows amino acid frequencies to be estimated, as well as allowing gamma distributed rates to be estimated).

The AIC_c calculation indicated that all models could be rejected in favour of the WAG+Γ model, in which rates of evolution are allowed to vary across sites according to a gamma distribution, but amino acid frequencies are not estimated from the data. For SOWHL tests using the JTT replacement matrix, the JTT+Γ model was used.

References

Hurvich CM, Tsai CL (1989) Regression and Time Series Model Selection in Small Samples. Biometrika 76(2): 297-307.

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8(3): 275-282.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9): 817-818.

Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53(5): 793-808.

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18(5): 691-699.