Influence of the background Markov model order

The order of the background Markov model influences the results. In mammalian genomes, nucleotide frequencies depend on the neighbouring nucleotides which can be seen most prominently at the level of CpG occurrences. To study the effect of the background model order on the results, we randomly selected sequences from the mouse genome and implanted motifs. Across all methods analysed in this study, the first order Markov model produces the best results (see Table 1). We therefore selected a first order Markov model throughout the analysis.

| Markov model order variations, m4r2, k=5 | AUC ROC | AUC PR 5%-Precision |
|----------------------------------------|---------|---------------------|
|                                       | M0      | M1      | M0      | M1      | M0      | M1      |
| D2                                    | 0.51    | 0.51    | 0.50    | 0.50    | 0.52    | 0.52    |
| D2z                                   | 0.57    | 0.67    | 0.54    | 0.62    | 0.57    | 0.67    |
| D2*                                   | 0.58    | 0.82    | 0.55    | 0.81    | 0.58    | 0.91    |
| N2*                                   | 0.61    | 0.90    | 0.58    | 0.89    | 0.63    | 0.94    |

Table 1: Influence of background Markov model order on pairwise scores. The data is obtained from randomly selected sequences from the mouse genome. Choosing a first order Markov model as background is always preferable.
**Influence of single sequence noise and repeats**

Alignment-free sequence comparison methods can easily be influenced by nucleotide composition and low complexity or repetitive sequences (see main text). Supplementary Figures 1 and 2 show that the N2 variants that use mismatches and the reverse complement as extended word neighbourhood are robust against a change of nucleotide distribution, they perform as would be expected.

![Influence of sequence composition on pairwise scores, uniform](image1)

![Influence of sequence composition on pairwise scores, AT-rich](image2)

Figure 1: Influence of single sequences on pairwise scores on the N2 variants, uniform nucleotide distribution. See main text for details.

Figure 2: Influence of single sequences on pairwise scores on the N2 variants, AT-rich distribution. See main text for details.

Mammalian genome sequences contain to a large degree repetitive sequences. We studied the influence of repeats on pairwise scores by randomly selecting sequences from the mouse genome and implanting 4 words of length 6 each twice into the sequences. Table 2 shows the comparison between repeat-masked and non-masked sequences. Repeat-masking is always recommended. N2 is most robust against repeats, whereas D2z is strongly influenced by repeats.
## Supplementary information

'Estimation of Pairwise Sequence Similarity of Mammalian Enhancers with Word Neighbourhood Counts'  
by J.Göke, M.H.Schulz, J.Lasserre and M.Vingron

| Repeat masked: | AUC ROC | AUC PR | Precision at 5%recall |
|----------------|---------|--------|-----------------------|
|                | N       | Y      | N                    | Y                      |
| \( D_2 \)      | 0.51    | 0.50   | 0.50                 | 0.49                   | 0.52                   | 0.52                   |
| \( D_2^z \)    | 0.67    | 0.71   | 0.62                 | 0.69                   | 0.67                   | 0.82                   |
| \( D_2^* \)    | 0.82    | 0.81   | 0.81                 | 0.83                   | 0.91                   | **0.97**               |
| \( N_2^* \)    | **0.90**| **0.90**| **0.89**            | **0.89**              | **0.94**              | **0.97**              |

Table 2: Comparison of the different methods. Values are averages over 25 simulations, numbers in bold indicate the best performing method for the given setting. Repeat-masking improves the results, \( N_2 \) produced the best results on unmasked sequences.
Supplementary information

‘Estimation of Pairwise Sequence Similarity of Mammalian Enhancers with Word Neighbourhood Counts’

by J. Göke, M. H. Schulz, J. Lasserre and M. Vingron

Figure 3: Influence of the mismatch weights $a_w$ on the pairwise scores of simulated sequences. The combination of $k = 6$ and $k = 5$ with $a_w = 1.0$ produced the best results. For $k = 4$, smaller mismatch weights gave better results.

![Graph showing influence of mismatch weights](image)

Figure 4: Precision-Recall plot for different choices of $a_w$. Shown is the Precision-Recall plot for sequences with words implanted from the neighbourhood of a randomly chosen motif. With small choices of $a_w$ close to 0.0 the mismatch variant $N2^{mm}$ performs similar to $N2^*$. Larger values of $a_w$ give better results.

![Graph showing precision-recall plot](image)

Choice of parameters on simulated sequences

Parameters that can be changed besides the Markov model order (see above) are the size of the k-mers $k$ and the weight for the word neighbourhood counts $a_w$. Supplementary Figure 4 shows that the influence of increasing $a_w$ is not linear in the increase in performance. For example, $a_w = 0.1$ (AUC PR 0.56) is closer in performance to $a_w = 1.0$ (AUC PR 0.57) than to $a_w = 0.01$ (AUC PR 0.54). A higher value of $a_w$ seems to give better performance with larger $k$ (see supplementary Figure 3). This is confirmed on real enhancer sequences (see supplementary Figure 5). In the main text we use $k = 6$ as size of the k-mers. This value was chosen since $k = 6$ gives the best results for real enhancer sequences (see below), we maintained the same set of parameters throughout this study to maintain consistency. In order to give a more complete picture we provide results using $k = 5$ on simulated sequences (Tables 3, 4). For a full overview of parameters on embryonic enhancer sequences, see next paragraph.
### Table 3: Comparison of different methods on simulated sequences, words are implanted on both strands, $k = 5$.

| Motif setting | $D_2$ | $D_2^z$ | $D_2^*$ | $N_2^*$ | $N_2^{rc}$ |
|---------------|-------|---------|---------|---------|------------|
| m1r8 m4r2     | 0.59  | 0.67    | 0.69    | 0.75    | 0.75       |
| m1r8 m4r2     | 0.74  | 0.79    | 0.59    | 0.74    | 0.74       |
| m1r8 m4r2     | 0.73  | 0.58    | 0.74    | 0.58    | 0.58       |
| m1r8 m4r2     | 0.53  | 0.57    | 0.58    | 0.58    | 0.58       |
| m1r8 m4r2     | 0.69  | 0.68    | 0.75    | 0.74    | 0.69       |
| m1r8 m4r2     | 0.73  | 0.81    | 0.62    | 0.81    | 0.62       |
| m1r8 m4r2     | 0.54  | 0.56    | 0.52    | 0.56    | 0.56       |
| m1r8 m4r2     | 0.55  | 0.55    | 0.55    | 0.55    | 0.55       |
| m1r8 m4r2     | 0.56  | 0.56    | 0.52    | 0.55    | 0.56       |
| m1r8 m4r2     | 0.54  | 0.56    | 0.52    | 0.55    | 0.56       |
| m1r8 m4r2     | 0.57  | 0.57    | 0.52    | 0.57    | 0.57       |
| m1r8 m4r2     | 0.56  | 0.57    | 0.52    | 0.56    | 0.57       |

### Table 4: Comparison of different methods on simulated sequences, implanted words are sampled from the mismatch-neighbourhood, $k = 5$.

| Motif setting | $D_2$ | $D_2^z$ | $D_2^*$ | $N_2^*$ | $N_2^{mm(0.1)}$ | $N_2^{mm(1.0)}$ |
|---------------|-------|---------|---------|---------|-----------------|-----------------|
| m1r8 m4r2     | 0.51  | 0.55    | 0.56    | 0.52    | 0.63            | 0.65            |
| m1r8 m4r2     | 0.53  | 0.55    | 0.56    | 0.52    | 0.54            | 0.54            |
| m1r8 m4r2     | 0.53  | 0.55    | 0.56    | 0.52    | 0.55            | 0.55            |
| m1r8 m4r2     | 0.53  | 0.55    | 0.56    | 0.52    | 0.55            | 0.55            |
Figure 5: Influence of the mismatch weights $a_w$ on the pairwise scores of developmental enhancers. In all data sets, the combination of $k = 6$ and $a_w = 1.0$ produced the best results. For $k = 4$, smaller mismatch weights gave better results, probably because words in the neighbourhood are more likely to occur by chance. (A) Forebrain. (B) Midbrain. (C) Heart. (D) Limb.

**Choice of parameters on embryonic enhancers**

The parameters that can be chosen freely are the background model order, the size of the k-mers ($k$) and the weight for the words with mismatch ($a(w)$). In order to estimate the robustness of the methods to changes of parameters, we run all methods with different $k$ and $a(w)$ for all data sets. We randomly selected genomic sequences from the mouse genome of the same size as these enhancers as the negative control data set. For this analysis, we combined the positive and negative set and calculated all pairwise scores. The values are averaged over 20 samples. For every sample we randomly selected 500 sequences from the positive set and ranked all pairwise scores together with all pairwise scores from 500 randomly selected negative sequences.

Supplementary Figure 5 shows the influence of the mismatch weight parameter $a_w$ for $k = 4, 5, 6$. In all data sets, the combination of $k = 6$ and $a_w = 1.0$ produced the best results. For $k = 4$, smaller mismatch weights gave better results, probably because words in the neighbourhood are more likely to occur by chance.

Supplementary Figures 6 to 8 show the Precision-Recall curve for $k = 4, 5, 6$. The neighbourhood concept consistently improved the results across all tissues and choices of $k$. Counting words with mismatches gave improvements for $k > 4$, consistent with supplementary Figure 5 which shows that a small mismatch weight gave better results for small $k$. 
Figure 6: Precision-Recall curve for enhancers active during mouse development, $k = 4$. The plots show the precision average over 25 samples each time drawing 500 enhancer sequences (’positive’) and 500 unrelated genomic sequences of equal length as the enhancers (’negative’). Results are shown for enhancers active in forebrain (A), midbrain (B), heart (C) and limb (D).
Figure 7: $k = 5$: Precision-Recall curve for enhancers active during mouse development. The plots show the precision average over 25 samples each time drawing 500 enhancer sequences ('positive') and 500 unrelated genomic sequences of equal length as the enhancers ('negative'). Results are shown for enhancers active in forebrain (A), midbrain (B), heart (C) and limb (D).

Figure 8: $k = 6$: Precision-Recall curve for enhancers active during mouse development. The plots show the precision average over 25 samples each time drawing 500 enhancer sequences ('positive') and 500 unrelated genomic sequences of equal length as the enhancers ('negative'). Results are shown for enhancers active in midbrain (A) and heart (B).
Supplementary information

‘Estimation of Pairwise Sequence Similarity of Mammalian Enhancers with Word Neighbourhood Counts’
by J. Göke, M.H. Schulz, J. Lasserre and M. Vingron

Figure 9: Precision-Recall curve for mouse enhancers. Enhancers active in different tissues were used as the background set.

Tissue-specific enhancers

These plots show the performance of pairwise comparison with alignment-free methods for enhancers active in the same tissue versus enhancers active in different tissues. The performance is reduced compared to randomly selected genomic sequences. Nevertheless, enrichers active in the same tissue have higher pairwise scores. Enhancers obtained from embryonic heart tissue (supplementary Figure 9) can not be distinguished from other embryonic tissues by $N_2$, $D_{2z}$ or $D_{2}^*$. Midbrain and forebrain enhancers show greater similarities than heart enhancers, which leads to an area under the precision-recall curve which is less than 0.5. Heart enhancers show much weaker sequence conservation (Blow et al., 2010), we therefore expect that this set assembles a highly divergent set of enhancers with a limited number of shared transcription factor binding sites. Interestingly, the heart data set is the only data set were $D_2$ shows better performance than the other methods, this might be caused by repetitive sequences which escaped repeat-masking.