# XSTREAM P-Value Estimates

The following tables provide recommended settings and estimated P-values for TRs identified from **protein sequences**:

| **TR degeneracy** | Min Word Match (*i*) | Min Consensus Match (*I*) | Max Gaps (*g*) |
|---|---|---|---|
| High (H) | 0.7 | 0.7 | 3 |
| Moderate (M)* | 0.7 | 0.8 | 3 |
| None (N) | 1 | 1 | 0 |

| **TR Significance** | Min Copy No. (*e*) | Min Period (*m*) | Min Domain Length (*minD*) | P-value (N) | P-value (M) | P-value (H) |
|---|---|---|---|---|---|---|
| Very High | 3 | 3 | 20 | $<10^{-5}$ | $<10^{-4}$ | $<10^{-4}$ |
| High* | 2 | 2 | 10 | $<10^{-4}$ | <0.02 | <0.02 |
| Moderate | 2 | 1 | 5 | <0.1 | <0.1 | <0.1 |

*=default settings

P-value cutoffs are color-coded according to the TR degeneracy level that they represent. We calculated P-value cutoffs for TR significance by first randomly shuffling characters within every protein sequence from the large and diverse NCBI non-redundant (NR) protein dataset (~7M sequences, downloaded April 2008) XSTREAM was run on all random sequences using the three different colored degeneracy settings shown above (High, Moderate, None). P-values were estimated by taking the number of protein sequences identified by XSTREAM for different ranges of TR periods and copy numbers, and dividing that count by the total number of randomized proteins in the NR dataset. This analysis was repeated for each parameter set using two independently randomized versions of NR, and the results were completely reproducible ($R^2 \sim 1$).