Comparison of sequence masking algorithms and the detection of biased protein sequence regions – Supplement

David P. Kreil[*†] Christos A. Ouzounis[†]

Supplement

— Revision 1.2, compiled July 22, 2003 —

[*]Department of Genetics / Inference Group (Cavendish Laboratory), University of Cambridge
[†]Computational Genomics Group, The European Bioinformatics Institute, EMBL Outstation Cambridge, CB10 1SD, UK

Abstract

Motivation Separation of protein sequence regions according to their local information complexity and subsequent masking of low complexity regions has greatly enhanced the reliability of function prediction by sequence similarity. Comparison with alternative methods that focus on compositional sequence bias rather than information complexity measures have shown that removal of compositional bias yields at least as sensitive and much more specific results. Besides the application of sequence masking algorithms to sequence similarity searches, the study of the masked regions themselves is of great interest. Traditionally, however, these have been neglected despite evidence of their functional relevance.

Results Here we demonstrate that compositional bias seems to be a more effective measure for the detection of biologically meaningful signals. Typical results on proteins are compared to results for sequences that have been randomized in various ways, conserving composition and local correlations for individual proteins or the entire set. It is remarkable that low-complexity regions have the same form of distribution in proteins as in randomized sequences, and that the signal from randomized sequences with conserved local correlations and amino acid composition almost matches the signal from proteins. This is not the case for sequence bias, which hence seems to be a genuinely biological phenomenon in contrast to patches of low-complexity.

Availability Software in executable form is available on request from the authors.

Contact D. P. Kreil, Computational Genomics Group, The European Bioinformatics Institute, EMBL Outstation Cambridge, CB10 1SD, UK, e-mail: Kreil@ebi.ac.uk

Results for `seg` run with parameters as recommended for long non-globular regions

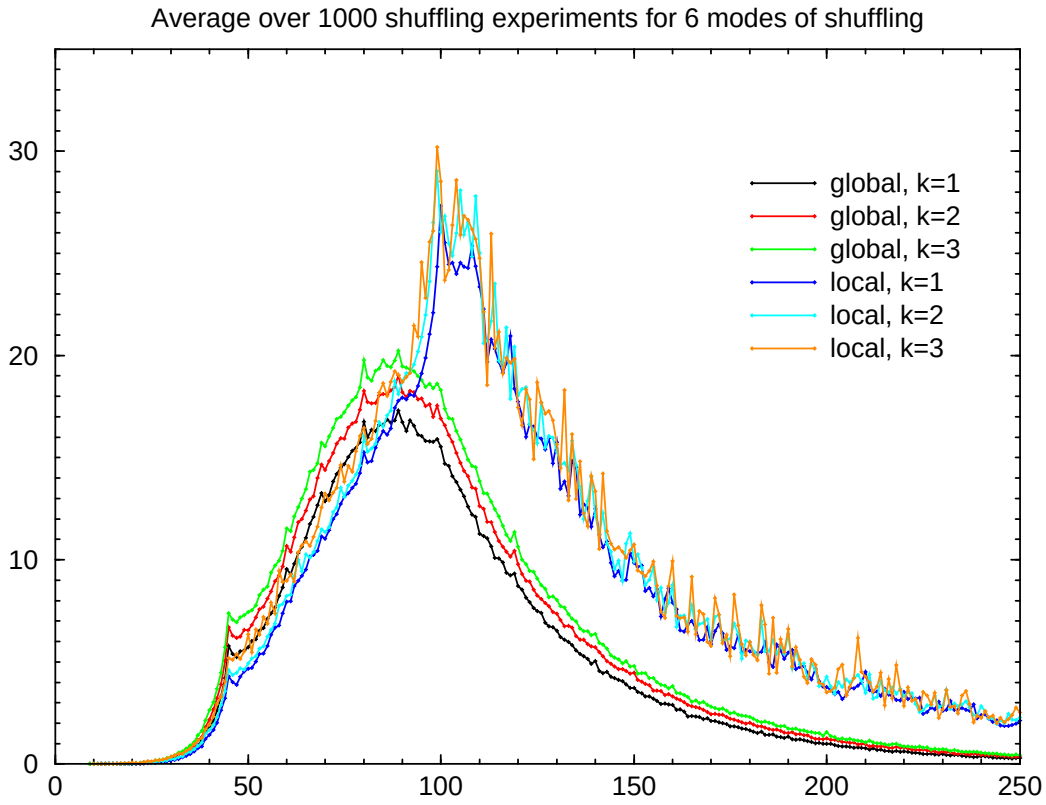

Figure 1: `seg` detected regions in shuffled sequences from *A. pernix*. The counts of regions reported by `seg` run with parameters recommended for detection of long non-globular regions (vertical axis) is plotted *vs* the lengths of these regions (horizontal axis) for different shuffling modes. See manuscript text for details.

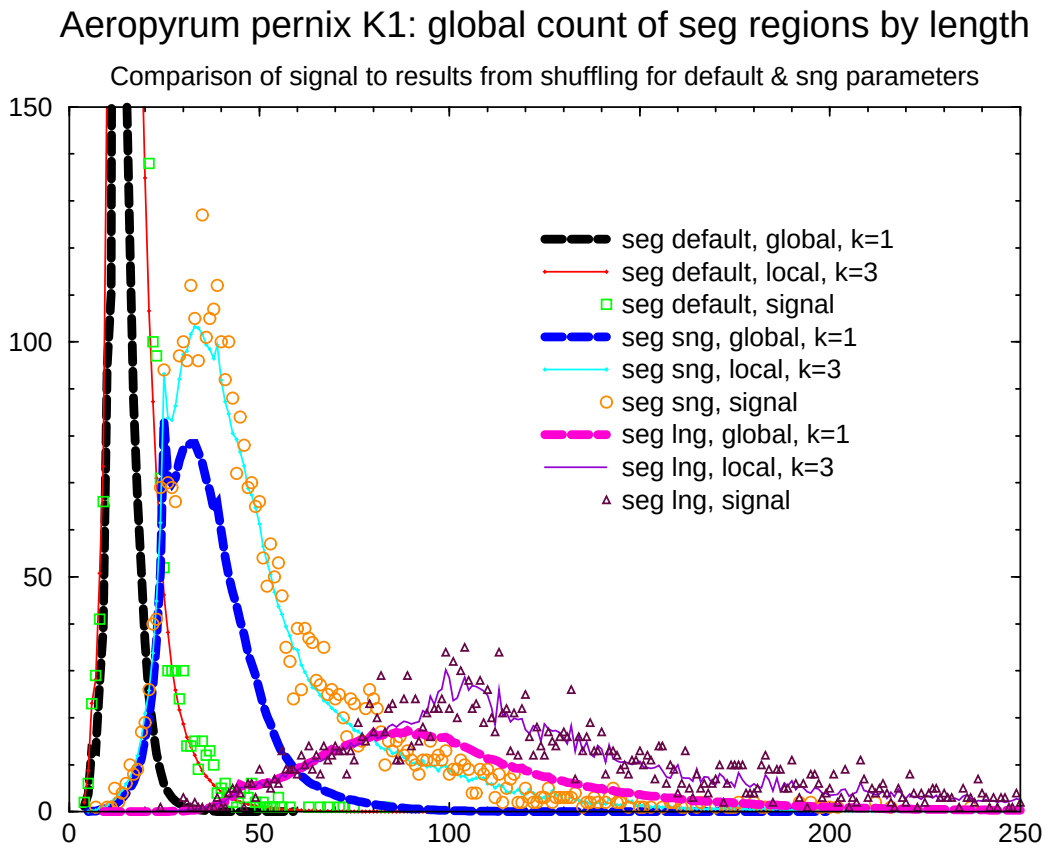

Figure 2: Signal to background characteristic of the `seg` method. This plot displays the count of low-complexity regions detected by `seg` in a sample organism, *A. pernix*, as a function of their lengths, for default parameters (green squares) and parameters recommended for the detection of small and long non-globular regions (orange circles/maroon triangles). In comparison, the respective counts for shuffled sequences from this organism are plotted (dashed/solid lines for globally/locally shuffled sequences).

Generation of randomized sequences: Influence of average amino acid composition on hit distributions

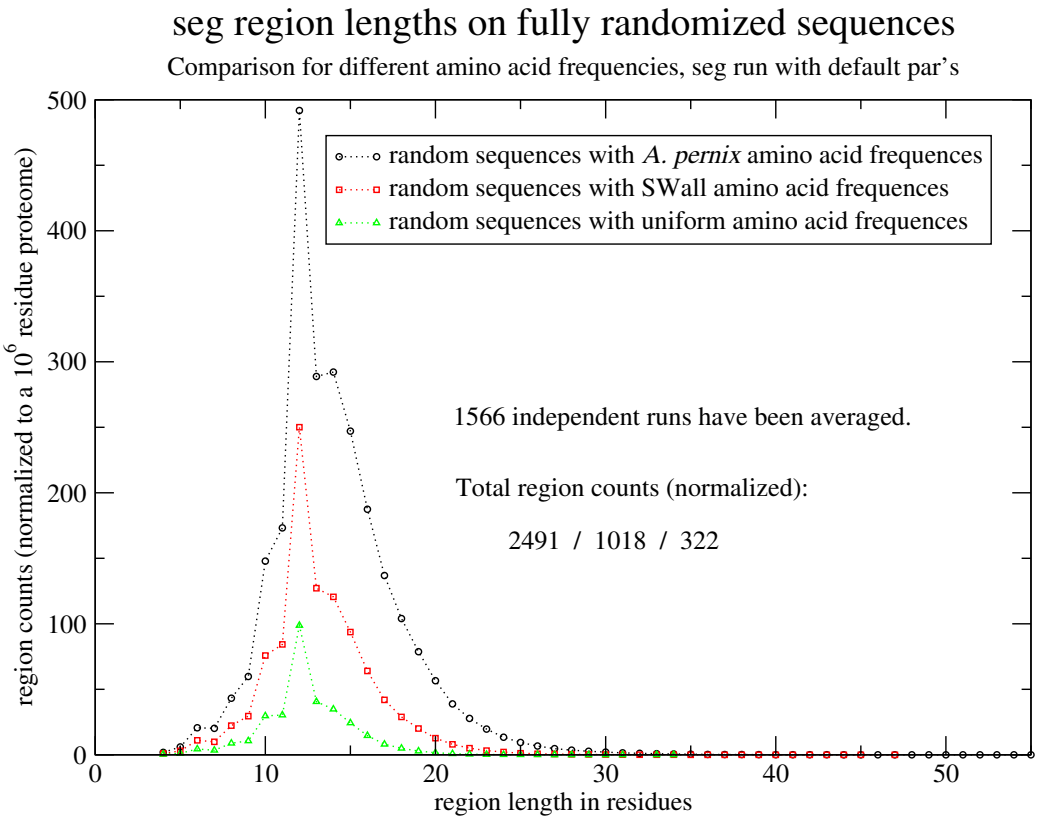

Figure 3: Influence of average amino acid composition on `seg` hit distribution. Random sequences of the same lengths as found in *A. pernix* have been generated according to given amino acid frequencies. The closer the amino acid distribution to `seg`'s model of uniform frequencies, the lower the total number of hits. The number of residues in detected regions also decreases: 3.5% / 1.3% / 0.40%. All counts have been normalized to a proteome length of one million residues to allow comparison with data sets of different sizes.

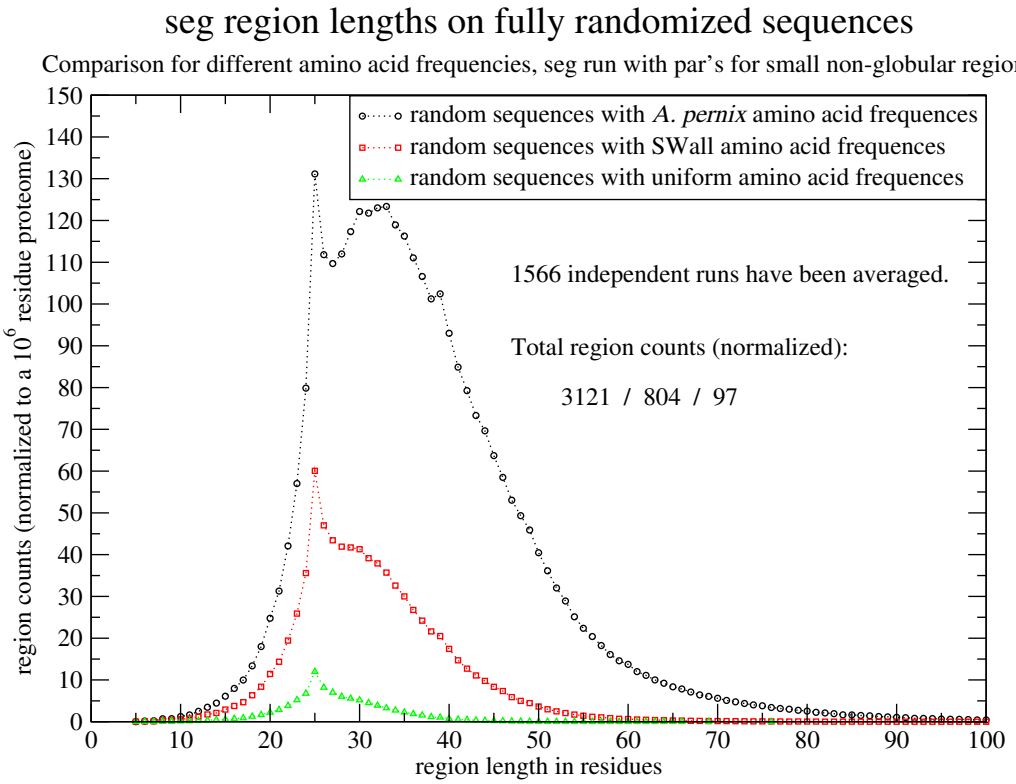

Figure 4: Influence of average amino acid composition on the distribution of `seg` detected small non-globular regions. Random sequences of the same lengths as found in *A. pernix* have been generated according to given amino acid frequencies. The closer the amino acid distribution to `seg`'s model of uniform frequencies, the lower the total number of hits. The number of residues in detected regions also decreases: 11.6% / 2.52% / 2.66%. One observes that this effect is even stronger with larger window length (compare Figs 3 and 5). All counts have been normalized to a proteome length of one million residues to allow comparison with data sets of different sizes.

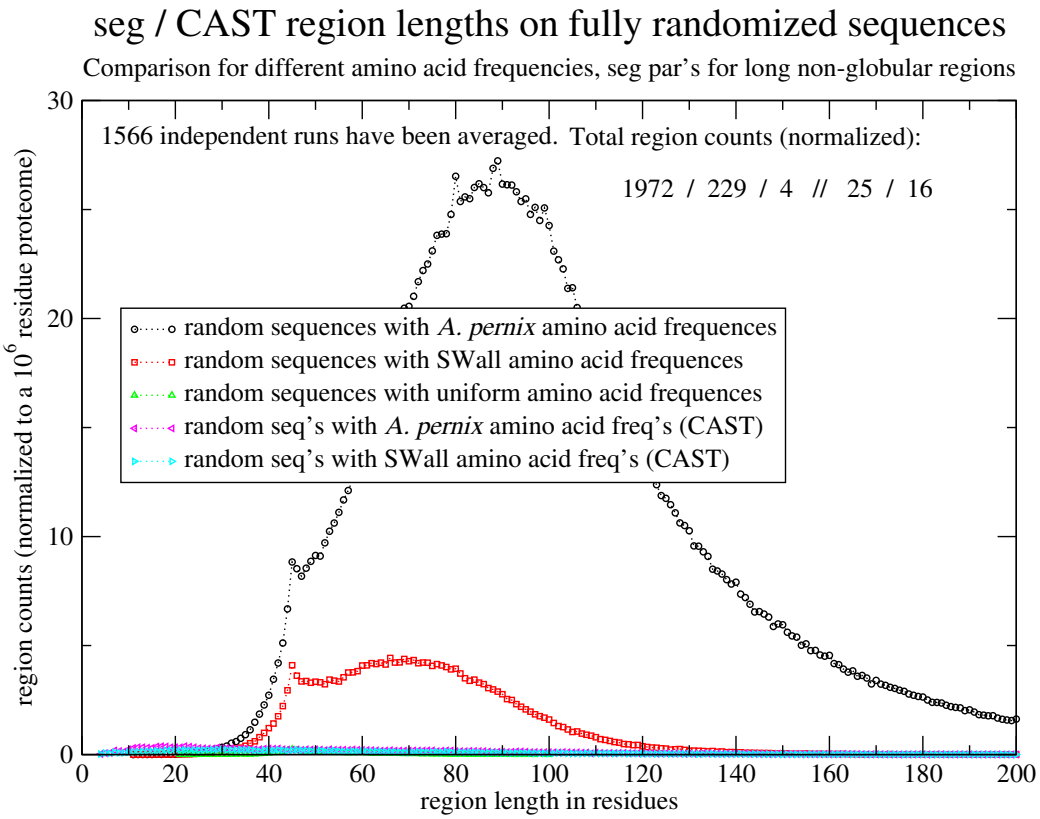

Figure 5: Influence of average amino acid composition on the distribution of `seg` detected long non-globular regions, and comparison to `CAST`. Random sequences of the same lengths as found in *A. pernix* have been generated according to given amino acid frequencies. The closer the amino acid distribution to `seg`'s model of uniform frequencies, the lower the total number of hits. The number of residues in detected regions also decreases: 20.2% / 1.67% / 0.020%. One observes that this effect is even stronger with larger window length (compare Figs 3 and 4). Counts for `CAST` are consistently low. See Fig. 6 for a region of the plot focusing on small count values. All counts have been normalized to a proteome length of one million residues to allow comparison with data sets of different sizes.

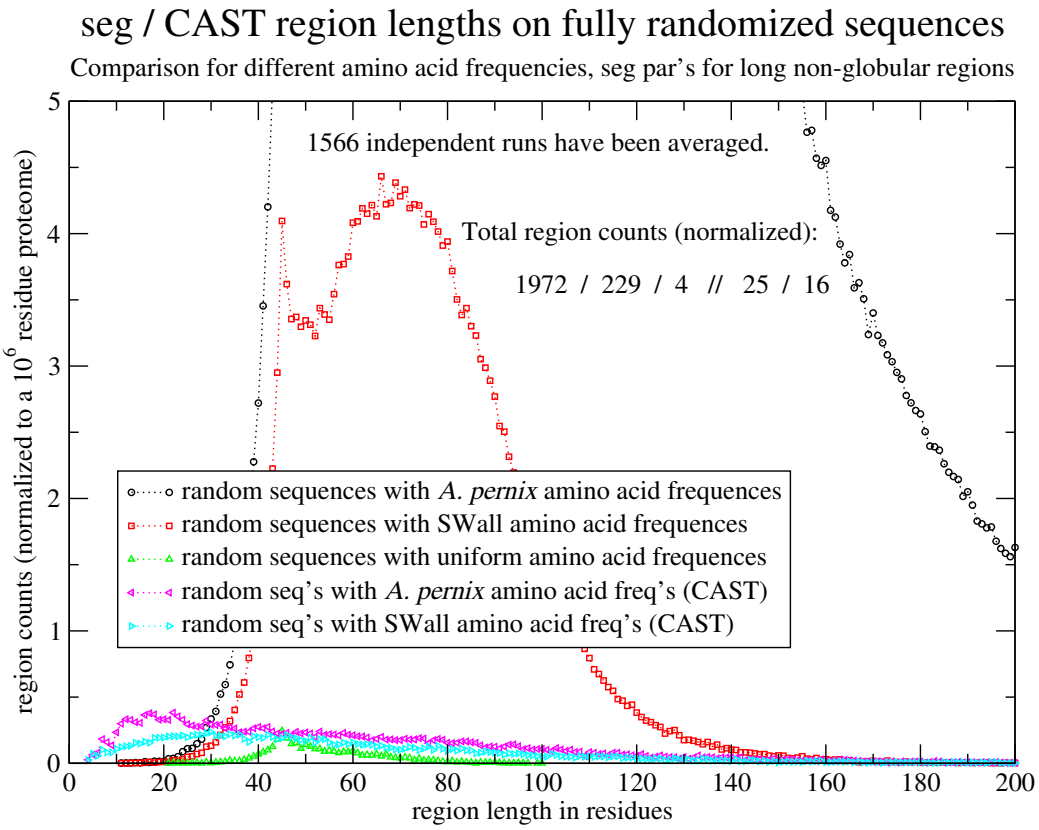

Figure 6: Comparison of the distribution of `seg` detected long non-globular regions to `CAST` detected regions, closeup. Random sequences of the same lengths as found in *A. pernix* have been generated according to given amino acid frequencies. Only for artificial sequences generated to conform to `seg`'s model of uniform frequencies is the number of `seg` detected long non-globular regions comparable to the number of `CAST` detected regions in random sequences of amino acid composition typical for real proteins. All counts have been normalized to a proteome length of one million residues to allow comparison with data sets of different sizes.