BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 16 2004, pages 2867–2868
doi:10.1093/bioinformatics/bth303

SWAPSC: sliding window analysis procedure to detect selective constraints

Mario A. Fares

Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland

Received on April 15, 2004; revised on April 21, 2004; accepted on April 25, 2004
Advance Access publication May 6, 2004

ABSTRACT

Summary: Sliding-window analysis procedure to detect selective constraints (SWAPSC) is a software system to dissect the constraints on the evolution of protein-coding genes. The program estimates rates of nucleotide substitutions at specific codon regions in each branch of a phylogenetic tree. The program uses several sets of simulated sequence alignments to estimate the probability of synonymous and non-synonymous nucleotide substitutions. Thereafter, a statistical analysis is conducted to determine the optimum window size to detect selective constraints. Finally, the optimum window size is slid along the real alignment and a test for significance of the estimated number of synonymous and non-synonymous nucleotide substitutions in each sliding step is conducted. A number of friendly useful output files is generated.

Availability: SWAPSC is available at http//www.may.ie/academic/biology/staff/mfmolecevolandbioinf.shtml. Distribution versions for both Linux and Windows operating systems are available, including manual and example files.

Contact: mario.fares@may.ie

Different structural and functional protein domains are likely to be subject to distinct selective constraints. Co-evolution between different codon sites within these domains questions the use of a single codon as the unit of selection, as previously demonstrated in several studies (Hughes and Nei, 1988; Marín *et al*., 2001).

Region-specific selective constraints can be determined by comparing the expected to the observed numbers of nucleotide substitutions. When the phylogeny is provided, the same approach can be applied to specific branches of the tree. SWAPSC enables precise analyses of selective constraints for each region of the sequence and branch of the tree. Briefly, SWAPSC estimates the average number of non-synonymous (θ_N) and synonymous (θ_S) nucleotide substitutions by the model of Li (1993) using simulated alignments. The probabilities of θ_N and θ_S are then calculated assuming a binomial distribution for each rate of substitution. These probabilities are used thereafter to obtain the optimum window size. Unlike other studies that use a random window (Tajima, 1991; Hughes and Nei, 1989), SWAPSC performs a statistical test to define the appropriate window size (Fares *et al*., 2002). Optimization of the window size is possible sliding a window of a defined size (e.g. 1–20 codons) along the simulated sequence alignment and testing the probabilities of synonymous (K_s) and non-synonymous (K_a) substitutions under a Poisson distribution. That window size showing the 5% lower $P(K_a)$ value >0.05 is selected as the appropriate one.

SWAPSC slides the optimum window along the real sequence alignment and estimates the probability of K_a and K_s comparing each sequence with its ancestor inferred by maximum parsimony. Due to that K_a and K_s are tested for significance, several hypotheses regarding the evolution of a specific branch of the tree and region of the alignment can be tested. Hence, SWAPSC can detect regions that are under adaptive evolution, or accelerated fixation rates of non-synonymous substitutions, or saturation of synonymous sites or hot spots.

The emphasis in SWAPSC has been focused in four main points: accuracy of results, automatic performance, accessibility and exhaustive screening of an unlimited alignment size. Files required and generated by SWAPSC are depicted in Figure 1A. An input multiple-alignment of coding sequences in PHYLIP format, standard to many different programs (e.g. PHYLIP, PAML, etc.), is required. Several sets of simulated sequence alignments in PHYLIP format and the tree in newick format have also to be provided. Then the user has two options: (1) fix the window size, which is not recommended unless biological information supports that option and (2) infer the appropriate window size.

The program generates independent output files in addition to the main output file to help the user to deal with the huge amount of information obtained. These files are: (1) an Excel file containing the information for each region and branch; (2) a file to visualize branches and locate constraints in an easy way using TREEVIEW program; and (3) a file with the amino acid replacements in each branch of the tree. Figure1B exemplifies the information that can be obtained.

The performance of the algorithm has been examined by several case studies (Fares *et al*., 2002; Lynn *et al*., 2004). To obtain more statistically robust results, the user is advised to: (1) use large datasets of sequence alignments; (2) use long

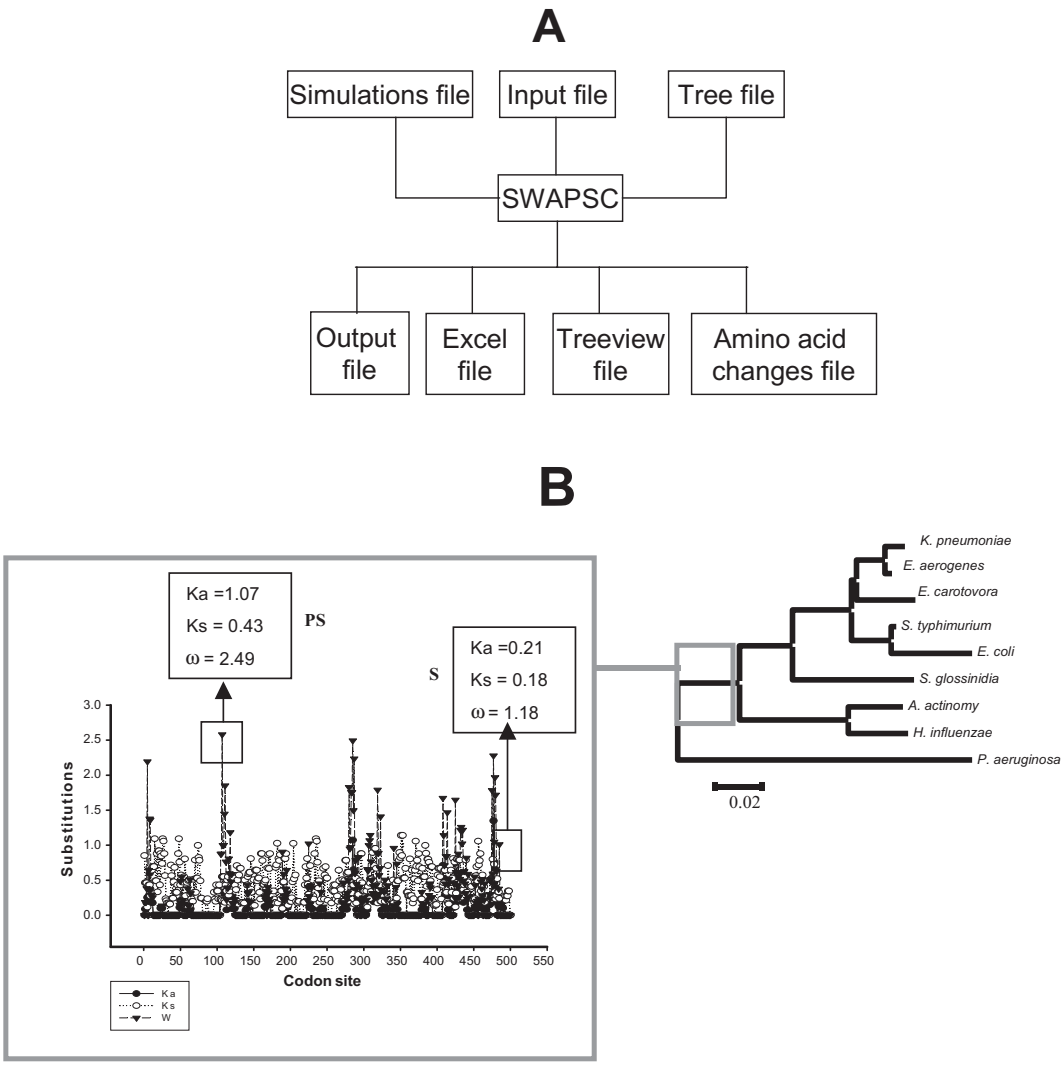

Fig. 1. Information generated by SWAPSC. (**A**) Files required or generated by SWAPSC and (**B**) selective constraints operating in one branch of the tree. Saturation of synonymous sites (S), positive selection (PS), non-synonymous (K_a) and synonymous (K_s) substitutions are indicated.

sequence alignments (200–300 codon sites); and (3) multiple alignments should be reliable.

The current version allows the use of Li's model, However, I will upgrade my software periodically to introduce more Kimura-based models.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr Avril Coghland for careful reading of the manuscript and the beta testers of SWAPSC for identifying bugs.

REFERENCES

Fares,M.A., Elena,S.F., Ortiz,J., Moya,A. and Barrio,E. (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.*, **55**, 509–521.

Hughes,A.L. and Nei,M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Hughes,A.L. and Nei,M. (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl Acad. Sci., USA*, **86**, 958–962.

Li,W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.

Lynn,D.J., Lloyd,A.T., Fares,M.A. and O'Farrelly,C. (2004) Evidence of positively selected sites in mammalian α-defensins. *Mol. Biol. Evol.*, **21**, 819–827.

Marín,I., Fares,M.A., González-Candelas,F., Barrio, E. and Moya,A. (2001) Detecting changes in the functional constraints of paralogous genes. *J. Mol. Evol.*, **52**, 17–28.

Tajima,F. (1991) Determination of window size for analysing DNA sequences. *J. Mol. Evol.*, **33**, 470–473.