The comparison of program sources using the sequence alignment of tokens

© A.V. Dubanov

Bauman Moscow State Technical University, Moscow, 105005, Russia

Borrowing detection is a very actual problem now. In this work, one of the known algorithms of the biopolymer sequence alignment was modified to make it possible to compare program sources and detect similar snippets in these sources. The input data of this algorithm are the source codes treated as the sequences of symbols. The set of lexical domains correspond to the alphabet of symbols making up these sequences. The algorithm was implemented and demonstrated with some code fragments written in Scheme language. The perspectives and restrictions of the algorithm application are also discussed.

Keywords: *borrowing of code sequence alignment, the largest common subsequence, dynamic programming, lexical analysis, functional programming.*

REFERENCES

[1] Burrows S., Tahaghoghi S.M.M., Zobel J. Efficient plagiarism detection for large code repositories. *Softw. Pract. Exper*, 2007, no. 37(2), 151–175.

[2] MOSS (Measure of Software Similarity). Available at: http://theoy.stanford.edu/~aiken/moss/ (accessed on 02.10.2014).

[3] Agrawal A., Huang X. Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2011, no. 8(1), pp. 194–205.

[4] Lewis J., Ossowski S., Hicks J., Errami M., Garner H.R. Text similarity: An alternative way to search MEDLINE. *Bioinformatics*, 2006, no. 22 (18), pp. 2298–2304.

[5] Durbin R., Eddy S.R., Krogh A., Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998, 336 p.

Dubanov A.V., Ph. D., assoc. professor of the Computer Science and Technologies Department of Bauman Moscow State Technical University. Specializes in the application of the computational methods and software development for biomedical research. He is the author of 20 publications. e-mail: qrcs@mail.ru.