Mathematical Problems of Computer Science 23, 2004, 134–143.

A Dynamic Programming Approach for Computing Similarity of the Protein Sequences Based on Continuous Functions Comparison

Robert K. Gevorgyan

Yerevan State University, Dep. of Applied Mathematics and Informatics.
e-mail robertg@ysu.am

Abstract

This paper introduces a dynamic programming approach for computing "continuous" similarity of the two protein sequences. The discrete dynamic programming method considers items of each comparable sequence independently; meantime there is a strong interrelation between them. To overcome this disadvantage a "continuous" sequence comparison method is developed. Particularly, a certain continuous function is correlated to each comparable protein sequence, and then the comparison is made between those functions. Through compressions and expansions the comparable functions are brought to the most similar representation in the meaning of a certain similarity function. By this approach the sequence comparison problem is reduced to a functional maximization problem, which is numerically solved using dynamic programming method. Finally some practical results are presented with the application of described method.

References

[1] Alexeev V.M., Tikhomirov V.M. and Fomin S.V., *Optimal Control.*, Nauka, 1979.

[2] Alstchul S.F., Glish W., Miller W., Myers E.W. and Lipman D.J. *Basic local alignment search tool.* J. Mol. Biol. 215, 403-410, 1990.

[3] Bairoch A. and Apweiler R. *The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.* Nucleic Acid Res., 27, 49-54, 1999.

[4] Baldi P. and Brunak S. *Bioinformatics, The Machine Learning Apprach.* MIT Press, 2001.

[5] Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M. and Sonnhammer L.L.E. *The Pfam protein families database.* Nucleic Acids Research, vol. 30, no. 1, 276-280, 2002.

[6] Bellman R. *Dynamic Programming.* Prinston Univ. Press, 1957.

[7] Durbin R. Eddy S.R., Krogh A., Mitchison G. *Biological Sequence analysis.* Cambridge University Press, 1998.

[8] Eddy S.R. *Profile hidden Markov models.* Bioinformatics, vol. 14, no. 9, 755-763, 1998.

[9] Gevorgyan R.K., *A Continuous Method for Evaluating Dissimilarity of the Protein Sequences.* Proceedings of the ISAAC Conference on Analysis, Yerevan, Armenia, 29-40, 2004.

[10] Gusfield D. *Algorithms on Strings, Trees, and Sequences.* Cambridge University Press, 1997.

[11] Heymann S., Gabrielyan O. R., Ghazaryan H. and others. *Towards a Metrical Space of Biological Sequences.* Proceedings of the ISAAC Conference on Analysis, Yerevan, Armenia, 1-18, 2004.

[12] Horst R., Pardalos P.M. and Thoai N.V. *Introduction to global optimization.* Kluwer Academic Publishers, 1995.

[13] Kantorovich L.V. and Rubinstein G.S. *On a function space and certain extremum problem.* Dokl.Akad. Nauk SSSR, N5, 115, 1058-1061, 1957.

[14] Levenstein V.I. *Binary codes capable of correcting insertions and reversals.* Sov. Phys. Dokl., 10:707-710, 1966.

[15] Mikhalevich V.S. *Sequential optimization algorithms and their applications.*, Kibernetika, N 1, 2, 1965.

[16] Moiseev N. N. *Calculus of approximations in the theory of optimal tasks.* Nauka, Moscow, 1971.

[17] Needelman S.B. and Wunsch C.D. *A general method applicable to the search for similarities in the amino acid sequences of two proteins.* J. Mol. Biol., 48, 443-453, 1970.

[18] Pearson W.R. and Lipman D.J. *Improved tools for biological sequence comparison.* Proc. Nat. Acad Sci. USA, 85, 2444-2448, 1988.

[19] Rabiner R. and Juang B.-H. *Fundamentals of speech recognition.* Prentice Hall PTR Englewood Cliffs, New Jersey 07632, 1993.

[20] Sankoff D. and Kruskal J.B. *Time Warps, String Edits and Macromolecules.* CSLI Publications, 1997, ISBN 1-57586-217-4.

[21] Setubal J.C. and Meidanis J. *Introduction to computational molecular biology.* PWS Publishing company, 1997.

[22] 21. Smith T.F. and Waterman M.S. *Identification of common molecular subsequences.* Journal of Molecular Biology, 147: 195-197, 1981.

[23] 22. Wasserstein L.N. *Markov processes over denumerable products of spaces describing large systems of automata.* Problems of Information Transission 5, 47-52, 1969.

[24] 23. Waterman M. *Introduction to computational biology.* Chapman and Hall, 1995.

Սպիտակուցային հաջորդականությունների՝ անընդհատ ֆունկցիաների վրա հիմնված նմանության հաշվարկումը դինամիկ ծրագրավորման մեթոդով:

Ռ. Կ. Գևորգյան

Ամփոփում

Այս հոդվածում ներկայացված է կենսաբանական հաջորդականությունների համեմատման մի մեթոդ: Հայտնի դիսկրետ դինամիկ ծրագրավորման մոթողը յուրաքանչյուր համեմտվող հաջորդականությունների անդամները դիտարկում է իրարից անկախ, այնինչ դրանց միջև կան որոշակի կապեր: Այդ թերությունը հաղթահարելու համար դիտարկվում է հաջորդականությունների համեմատման մի 'անընդհատ' մեթոդ: Այն է՝ յուրաքանչյուր համեմատվող հաջորդականությանը համապատասխանության մեջ է դրվում մի անընդհատ ֆունկցիա և ապա համեմատությունը կատարվում է այդ ֆունկցիաների միջև: Սեղմումների և ձգումների միջոցով համեմատվող ֆուկցիաները բերվում են ամենանման տեսքի՝ տրված նմանության ֆունկցիայի իմաստով: Այս մոտեցման միջոցով խնդիրը բերվում է ֆունկցիոնալի օպտիմիզացիայի խնդրի, որը թվապես լուծելու համար ներկայացված է մի դինամիկ ծրագրավորման մեթոդ: Աշխատանքում բերված են նաև որոշ պրակտիկ արդյունքներ՝ ներկայացված մեթոդի կիրառմամբ: