Simultaneous Identification of Conserved Regions and Inversions through Global Optimization

Alair Pereira do Lago
Computer Science Department
Universidade de São Paulo
alair@ime.usp.br

Abstract

We propose a new method for the comparative analysis of two long DNA sequences. Our method is the first that takes into account biological events like local inversions when searching for highly conserved regions.

Our method introduces new ideas to the usual concept of alignments, that detects events like insertions, deletions and mutations but not inversions. It proposes a weighted bipartite-matching approach to account for inversions. There is a polynomial solution to this approach, but the obtained solution may produce biologically unrealistic number of inversions. If cost for inversions is added, no polynomical solution is known.

We propose a model that supposes non-overlapping inversions that leads to a polynomial algorithm. This method has been successfully applied to public genomic data availabel for *Xylella fastidiosa* and *Pseudomonas aeruginosa*.

References

[1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* **215** (3), 403–410.

[2] Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research,* **10** (7), 950–958.

[3] Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O. & Salzberg, S. L. (1999) Alignment of whole genomes. *Nucleic Acids Research,* **27** (11), 2369–2376.

[4] Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M. & Frazer, K. A. (2000) Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Research,* **10** (9), 1304–1306.

[5] Hannenhalli, S. & Pevzner, P. A. (1995) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. In *ACM Symposium on Theory of Computing* pp. 178–189. Association for Computing Machinery.

[6] Koonin, E. V. (1999) The emerging paradigm and open problems in comparative genomics. *Bioinformatics,* **15** (4), 265–266. Editorial.

[7] Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S. & Dubchak, I. (2000) VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics,* **16** (11), 1046–1047.

[8] Muchnik, I., do Lago, A. P., Llaca, V., Linton, E., Kulikowski, C. A. & Messing, J. (2001). Assignment-like optimization on bipartite graphs with ordered nodes as an approach to the analysis of comparative genomic data. DIMACS Workshop on Whole Genome Comparison. http://dimacs.rutgers.edu/Workshops/WholeGenome/.

[9] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research,* **27** (1), 29–34.

[10] Robinson, G. d. B. (1938) On the representations of the symmetric group. *American Journal of Mathematics,* **60**, 745–760.

[11] Schöniger, M. & Waterman, M. S. (1992) A local algorithm for DNA sequence alignment with inversions. *Bull Math Biol,* **54** (4), 521–536.

[12] Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. & Miller, W. (2000) PipMaker–a web server for aligning two genomic DNA sequences. *Genome Research,* **10** (4), 577–586.

[13] Simpson *et al.* (2000) The genome sequence of the plant pathogen *Xylella fastidiosa.* The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature,* **406** (6792), 151–157.

[14] Stover *et al.* (2000) Complete genome sequence of *Pseudomonas* aeruginosa PA01, an opportunistic pathogen. *Nature,* **406** (6799), 959–964.

[15] Wagner, R. (1975) On the complexity of the extended string-to-string correction problem. In *Seventh ACM Symposium on the Theory of Computation* Association for Computing Machinery.

[16] Weiner, P. (1973) Linear pattern matching algorithms. In *14th Annual IEEE Symposium on Switching and Automata Theory (Univ. Iowa, Iowa City, Iowa, 1973).* IEEE Comput. Soc., Northridge, Calif. pp. 1–11.