Benedetto, Caglioti, and Loreto Reply: In [1] Khmelev *et al.* claim that [2] contains many misleading statements and that Markov chain approaches represent a more attractive technique than Lempel-Ziv (LZ)-based compression schemes.

The authors recall the results of the experiments reported in [3] where it is shown that Markov chain based methods outperform gzip based methods. First of all, it is important to remember that in [3] an LZ-based scheme (rarw) outperformed the Markov chain approach. The generic claim made in [1] that the Markov chain approach outperforms LZ-based schemes is then not so general. On the other hand, in our opinion, even the claim that Markov chain approaches outperform *gzip*-based schemes is not well supported by the experiments presented in [3].

In [3] the authors used a method which is slightly but crucially different from ours [2]. The authorship attribution of a text X is performed by comparing this file with one very long file per author to be used as the reference file (created by simply appending all the files of each single reference author in one single file) whose length is between 140 Kb and 6.9 Mb (average length 1.2 Mb). This procedure seems to us to be definitively wrong in the perspective of using the *gzip* compression scheme. *Gzip* has a sliding window of 32 Kb over which it looks for the longest matchings. Now, using as reference file a file longer than 32 Kb and appending to it the file X, only the last 32 Kb (which depend on the way the reference file has been created) are compared with the unknown file. Therefore most of the available information is lost. In our method [2] we use all the available files as reference files and we compare the unknown file X with all the other files. The indication of the authorship is given by the reference text closest to the unknown text.

In order to show how this slight difference in the procedure could bring drastically different results we have performed two sets of experiments. The first experiment concerns the classification of a corpus of newsgroup messages widely used for the comparison of different approaches [4]. The rate of success is 60% with the method of [3] and 85% with our method. The second experiment concerned the data presented in [2]. In this case our prescription allows for 93% (84/90) of success, while with the prescription adopted in [3] we get a success rate of 77% (69/90).

As for the usefulness of compression techniques for DNA analysis, in our opinion this is a very challenging field where it is almost impossible at present to say which method will be asymptotically successful. Just to give some examples, in [5] a method has been proposed to compress DNA sequences based on the backward search of approximate repeats. This allows one to define a distance between sequences useful to compare different genomes and build Phylogenetic trees.

As for our definition of remoteness [Eq. (1) in [2]] it is important to stress that relative entropy is a positive defined object. Since data compression schemes provide an estimate of the relative entropy it could happen that when the relative entropy is close to zero the approximation procedure could sort a slightly negative value. This phenomenon does not affect at all the validity of our results.

As for the objection concerning the coding chosen for our texts, one has to remember that a zipper "reads" the sequences of characters which one inputs to it, nothing more than this. The idea of comparing languages written with different alphabets cannot forget this simple statement. In order to compare languages written with different alphabets one should, for instance, consider texts written with the phonetic alphabet. This is the reason for not having included in our preliminary analysis of the language tree languages such as Chinese, Greek, Russian, etc.

As for the computational time we agree that data compression techniques can be slower than other procedures, but in our opinion the potentiality of the method is not only related to its speed.

Finally, we remark that only after the publication of our Letter did we become aware of the results by Loewenstern *et al.* [6] and of Khmelev [3] in which the idea of zipping a file B appended to a file A in order to define a remoteness between A and B had been previously stated.

Dario Benedetto and Emanuele Caglioti
Mathematics Department
"La Sapienza" University
P. le A. Moro 2
00185 Rome, Italy

Vittorio Loreto
Physics Department
"La Sapienza" University
P. le A. Moro 5
00185 Rome, Italy

Received 15 November 2002; published 27 February 2003
DOI: 10.1103/PhysRevLett.90.089804
PACS numbers: 89.70.+c, 01.20.+x, 05.20.–y, 05.45.Tp

[1] D.V. Khmelev and W.J. Teahan, preceding Comment, Phys. Rev. Lett. **90**, 089803 (2003).
[2] D. Benedetto *et al.*, Phys. Rev. Lett. **88**, 048702 (2002).
[3] See the appendix in O.V. Kukushkina, A. A. Polikarpov, and D.V. Khmelev, Prob. Peredachi Inf. **37**, No. 2, 96–108 (2000) [Probl. Inf. Transm. (Engl. Transl.) **37**, 172–184 (2001)].
[4] Available from http://www.ai.mit.edu/~jrennie/20Newsgroups.
[5] M. Li *et al.*, Bioinformatics **17**, 149–154 (2001).
[6] D. Loewenstern *et al.*, Rutgers University DIMACS Technical Report No. 95-04, 1995.

 © 2003 The American Physical Society