# On the characterization of DNA primary sequences by a new algorithm for similarity/diversity measures

R. Todeschini[a*], V. Consonni[a], A. Mauri[a], and D. Ballabio[b]

[a] Dept. of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1 – 20126 Milano (Italy)

[b] Dept. of Food Science and Technologies, University of Milan, via Celoria, 2 – 20133 Milano (Italy)

## Summary

In general, the proposed similarity/diversity measure appears as a new approach to sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements. In this paper, this methodology has been applied to one of the most interesting cases of sequential data, i.e. to DNA sequences. Then, a new distance measure derived from the partial ordering approach is proposed for evaluating the similarity/diversity among DNA sequences.

This distance – weighted standardized Hasse distance - is evaluated between pairs of Hasse matrices derived from the classical partial ordering rules. It can be naturally standardized, thus allowing to interpret these distances as absolute values (e.g. percentage) and deriving simple similarity and correlation indices.

DNA sequences taken from the first exons of the beta-globins for eight different species have been analyzed. Sensitivity analysis has been also performed, showing the high capability of this measure to take into account small modifications of the DNA sequences. Finally, a comparison with results obtained from literature is given, together with a comparison with matrix invariants derived from the Hasse matrix.