MERGE: a software package for generating a single data-base starting from EMBL and GenBank collections

M.Attimonelli, C.Lanave[1], S.Liuni and G.Pesole

Dipartimento di Biochimica e Biologia Molecolare and [1]Centro Studi Mitocondri e Metabolismo Energetico (CNR), University of Bari, Bari, Italy

Submitted August 11, 1987

MERGE is a software package which has been developed to compare EMBL and GenBank collections generating a new collection containing the GenBank entries along with EMBL entries not contained in the GenBank.

As shown in figure 1a, the following cathegories of sequences are identified:

a) EMBL-entries already present in GenBank collection, but updated in the latest EMBL release;

b) EMBL-entries not present in GenBank collection;

c) EMBL-entries containing partial GenBank sequences;

d) EMBL-entries which are part of GenBank entries.

Sequences of categories a),b) and c) are then transformed, using the program TRANSF, in GenBank data-format and inserted into our ACNUC database. The major problem was encountered during the transformation of the EMBL features table into the GenBank features and sites table, owing to the different syntax used

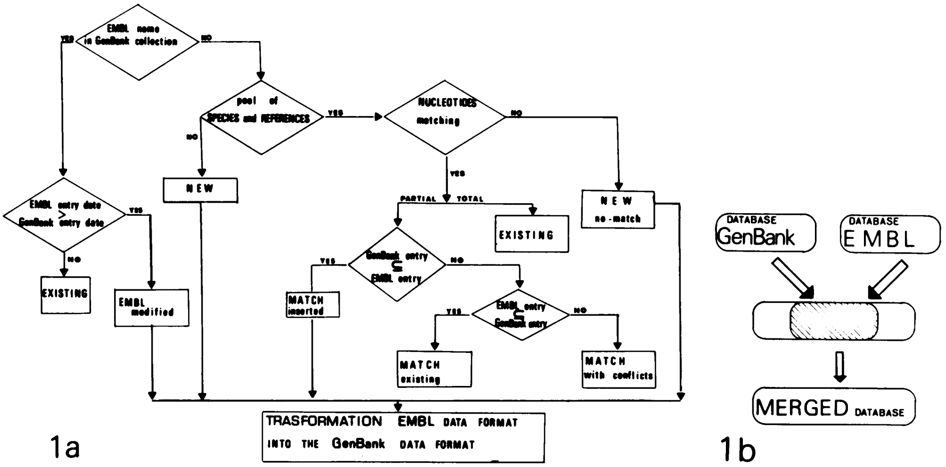
EMBL name in GenBank collection
YES
NO
pool of SPECIES and REFERENCES
YES
NO
NUCLEOTIDES matching
NO
YES
EMBL entry date > GenBank entry date
YES
NO
NEW
EXISTING
EMBL modified
PARTIAL
TOTAL
EXISTING
NEW no-match
GenBank entry ⊆ EMBL entry
YES
NO
MATCH inserted
EMBL entry ⊆ GenBank entry
YES
NO
MATCH existing
MATCH with conflicts
TRASFORMATION EMBL DATA FORMAT INTO THE GenBank DATA FORMAT
1a
DATABASE GenBank
DATABASE EMBL
MERGED DATABASE
1b

Fig. 1 - a) Flow chart of MERGE software
b) Merged collection obtained from the EMBL and the GenBank collections by removing the overlapping region.

in the two collections as well as to the incongruency encountered within each collection.
This software package has been developed in order to improve usage of two of our existing packages: ACNUC (1), for retrieval and extraction of the nucleic acid sequences; and GLORIA (submitted for publication) for analysis of extracted sequences. In fact, application of ACNUC and GLORIA on the merged database has proven to be extremely advantageous. In our opinion this work will be a great help to projects devoted to automatization of collection updating as well as in generation of specialized databases.

ACKNOWLEDGMENTS

We would like to thank prof.C.Saccone for her many helpful discussions and prof.J.Credico for the final english version. This work was partially supported by the "Progetto Finalizzato Ingegneria Genetica e Basi Molecolari delle Malattie Ereditarie" of the CNR - Italy, by the "Ministero della Pubblica Istruzione" - Italy and by the "Progetto Strategico Biotecnologie" of the CNR - Italy.

REFERENCES

1) Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and di Paola,G. (1985) CABIOS 1(3), 167-172