CABIOS

Vol.5. no.1. 1989
Pages 47 – 50

# SDSE: A software package to simulate the evolution of a pair of DNA sequences

José L.Oliver, Antonio Marín[1] and Juan-Ramón Medina[1]

## Abstract

*An algorithm to simulate DNA sequence evolution under a general stochastic model, including as particular cases all the previously used schemes of nucleotide substitution, is described. The simulation is carried out on finite, variable length, DNA sequences through a strict stochastic process, according to the particular substitution rates imposed by each scheme. Five FORTRAN programs, running on an IBM PC and compatibles, carry out all the tasks needed for the simulation. They are menu driven and interfaced to the system through a principal menu. All sequence data files used and generated by the SDSE package conform to the standard GenBank database format, thus allowing the use of any sequence retrieved from this databank, as well as the application of other packages to analyse, manipulate or retrieve simulated sequences.*

## Introduction

The evolutionary change of DNA occurs either by nucleotide substitution, or by deletion and insertion. We describe here the simulation of the first of these processes. The software allows the simulation of DNA evolution under a general stochastic model (F.Rodriguez, J.L.Oliver, A.Marin and J.R.Medina, in preparation) which includes as particular cases all the previously used schemes of nucleotide substitution (see Nei, 1987 for a review). Our general four hypothesis model (G4H) involves three assumptions with respect to the rates of nucleotide substitution; they are: (i) independent of sequence site, (ii) constant in time, and (iii) the same for the two sequences. A fourth assumption of the model is that base frequencies of the ancestral sequence are equilibrium frequencies, so that they remain unaltered during the evolutionary process. The software package described here also includes some complementary programs to carry out all the ancillary tasks needed by the simulation.

Simulation of DNA evolution is very useful when one wants to test, for example, the accuracy of a new index estimating the number of nucleotide substitutions or the evolutionary distance between two sequences, the power of a new DNA mathematical model, etc. Simulation has been extensively used in the study of DNA sequence evolution (Gojobori *et al.*, 1982; Tajima and Nei, 1984), but, to our knowledge, a general purpose program is not yet available.

## System and methods

The programs were developed on an IBM PC compatible computer running MS-DOS version 3.2. The minimum system configuration is 256K RAM (random access memory) and two disk drives. The programs are written in FORTRAN77 (Microsoft Fortran V3.30) and are available in source or machine (compiled) code. They will run on the IBM PC, IBM PC/XT, IBM PC/AT and compatibles. The current implementation does not require the 8087 numeric co-processor but, if present, it should not interfere. A compiled version exploiting the 8087 is also available.

## Algorithms

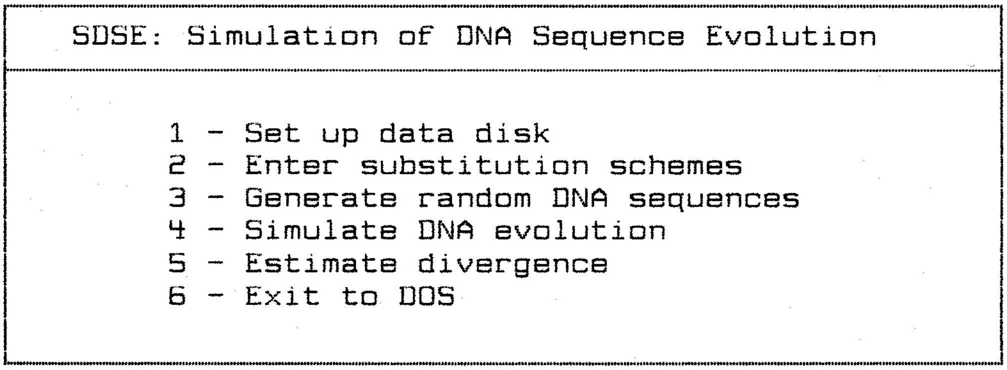
In the simulation of DNA evolution through nucleotide substitution, four main tasks can be identified: (i) the computation of nucleotide equilibrium frequencies for each substitution scheme; (ii) the generation of a random DNA sequence with those equilibrium nucleotide frequencies (to be used as ancestral sequence); (iii) the simulation of nucleotide changes on this ancestral sequence under different substitution schemes; and (iv) the estimation of the divergence accumulated by two 'evolved' sequences. All these tasks are readily carried out by the present software package. To facilitate its use, other utilities allowing both system set-up and entry data functions were added.

Equilibrium frequencies are computed by repeatedly squaring the corresponding substitution matrix, following Gojobori *et al.* (1982). Squaring is continued until all the rows in the matrix are equal. This task is carried out by the program XSCH.

Random DNA sequences of variable lengths and with nucleotide equilibrium frequencies are generated by the program XSEC. Frequencies are multiplied by the desired number of nucleotides in the sequence, and a vector containing the corresponding number of each nucleotide is constructed. A random sequence is then derived from this ordered vector by random extractions of nucleotides without replacement. This is carried out with the help of a random number generator whose seed is taken from the time provided by the system.

The simulation of nucleotide changes under different substitution schemes is conducted by the program XSIM as

*Unidad de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, and Dpto. de Genética, Facultad de Biologia, Universidad de Sevilla, Ap. 1095, E-41080 Seville, Spain*



```
SDSE: Simulation of DNA Sequence Evolution

        1 - Set up data disk
        2 - Enter substitution schemes
        3 - Generate random DNA sequences
        4 - Simulate DNA evolution
        5 - Estimate divergence
        6 - Exit to DOS
```

**Fig. 1.** Principal menu of SDSE package.

follows. A nucleotide site in the ancestral sequence was chosen at random. Any of the four nucleotides (A, T, C or G) could be located at this site. Let T be, for example, the nucleotide occupying this site. We divided the range 0−1 into four segments and used the substitution matrix to assign the length of each segment; the length of the first segment was made proportional to the probability that T does not change (T→T); the length of the three remaining segments were made proportional to the probabilities of T being substituted by A (T→A), C (T→C) or G (T→G). We then generated a random number in the range 0−1. The original nucleotide is not changed if this number pertains to the first segment, but was substituted by A, C or G if it pertains to the second, third or fourth segment, respectively. The same is realized with any nucleotide occupying the chosen site. A similar process is repeated until the sequence accumulates the desired number of changes, thus generating an 'evolved' sequence.

The divergence at nucleotide level between two 'evolved' sequences was determined by the program XDIV, which computes the nucleotide pair frequencies. The program can also take as input a matrix with the frequencies of different nucleotide pairs. The following methods were used to estimate nucleotide substitutions: Jukes−Cantor (1969) single parameter (JC) method, Kimura's (1980) two-parameter (K2P) method, Kimura's (1981) three-substitution-type (K3ST) method, Takahata and Kimura's (1981) (TK) method, Gojobori *et al.*'s (1982) method and Tajima and Nei's (1984) method.

## Implementation

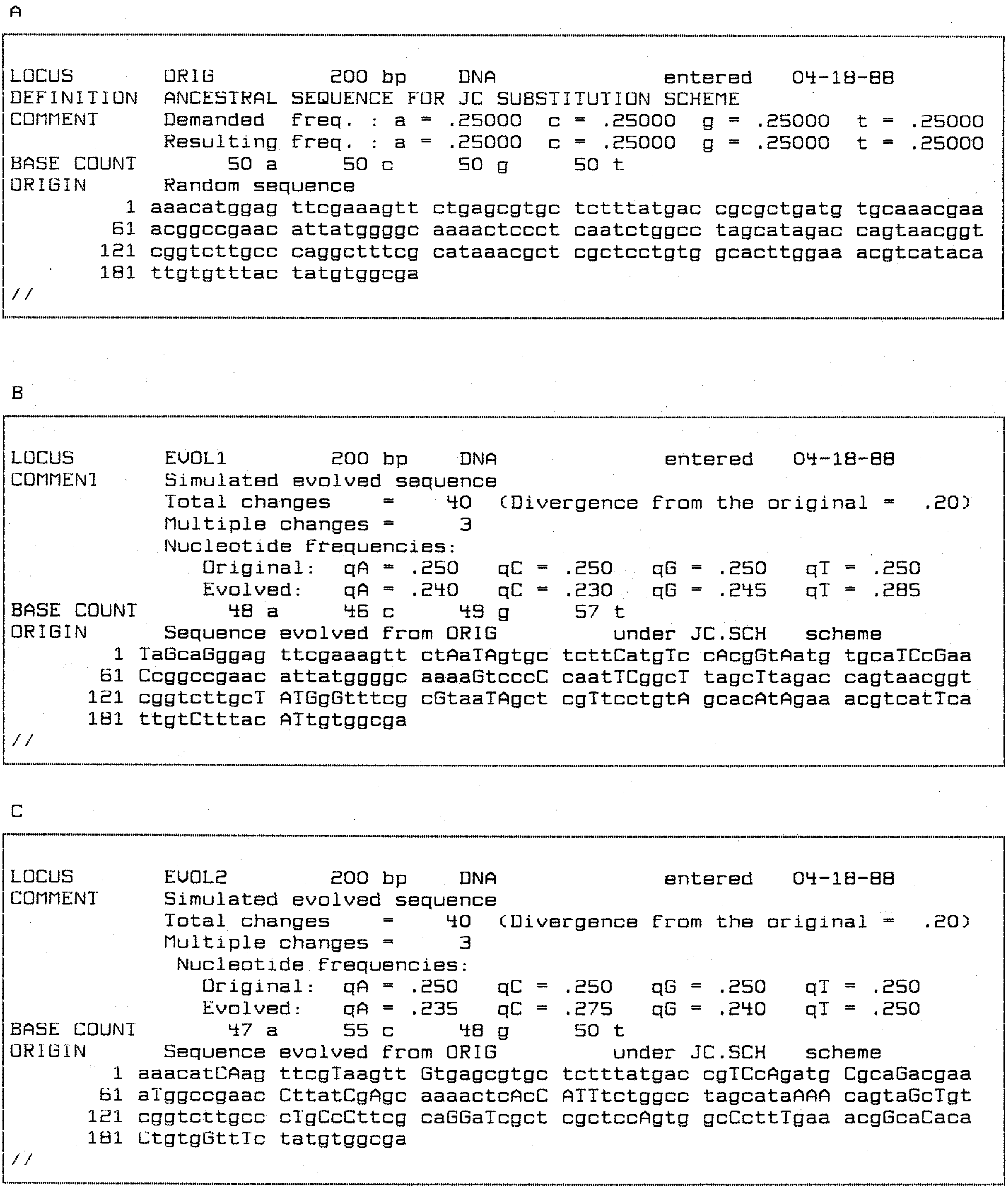
All the programs are menu driven and interfaced to the system through a principal menu (Figure 1). Detailed installation instructions are provided with the distribution disk.

The first choice is to set up the data disk drive. This program configures the package for a particular system. The disk arrives set up for the PC—that is, programs go in drive A, data on drive B, but other configurations can be intended, e.g. the programs can be copied to the hard disk and drive A used as data disk.

The second option allows the substitution schemes to be entered and computes the corresponding equilibrium nucleotide frequencies. Through an interactive process the user can fix the transition matrix corresponding to the chosen scheme, that is, fix the rates of nucleotide substitution per unit evolutionary time, assigning values to the different parameters used by the most common schemes: JC, K2P, K3ST, TK, GIN and TN. The user can also enter directly the 12 nucleotide substitution rates per unit evolutionary time, thus using a non-standard substitution scheme: OT. The edited scheme and the resulting equilibrium frequencies are first shown on the screen and then written to a file on the data disk, whose filename will be terminated in .SCH. Thus, one can have a file JC.SCH for the Jukes−Cantor substitution scheme, a file K2P.SCH for the Kimura-2P, etc.

By choosing option 3 of the principal menu, the user can generate random DNA sequences with desired nucleotide frequencies (Figure 2A). This utility thus allows the generation of an ancestral sequence with the equilibrium frequencies (computed by means of the option 2) corresponding to a particular substitution scheme. The maximum number of nucleotides is set to 3000, although it is an easy task to redimension the programs to accommodate larger sequences.

Since it seems to be convenient that the generated sequence has a standard format, it is structured in compliance with GenBank (Burks *et al.*, 1985) requirements. Thus, the program demands information for the directory entry of this database. The directory is the first section of the sequence file. The following directory lines are implemented by the program. The LOCUS line has the name of the sequence (a short unique name for the entry, chosen to suggest the sequence's DEFINITION), the number of nucleotides and the date the sequence was generated. The DEFINITION line includes a concise description of the sequence. In the COMMENT line the program includes both the demanded and the resulting nucleotide frequencies. The BASE COUNT line gives the different nucleotide counts. The ORIGIN line is the last line in the directory, immediately preceding the sequence itself, and describes the source of the



A

```
LOCUS         ORIG            200 bp     DNA                 entered    04-18-88
DEFINITION    ANCESTRAL SEQUENCE FOR JC SUBSTITUTION SCHEME
COMMENT       Demanded  freq. : a = .25000  c = .25000  g = .25000  t = .25000
              Resulting freq. : a = .25000  c = .25000  g = .25000  t = .25000
BASE COUNT        50 a      50 c      50 g      50 t
ORIGIN        Random sequence
          1 aaacatggag ttcgaaagtt ctgagcgtgc tctttatgac cgcgctgatg tgcaaacgaa
         61 acggccgaac attatggggc aaaactccct caatctggcc tagcatagac cagtaacggt
        121 cggtcttgcc caggctttcg cataaacgct cgctcctgtg gcacttggaa acgtcataca
        181 ttgtgtttac tatgtggcga
//
```

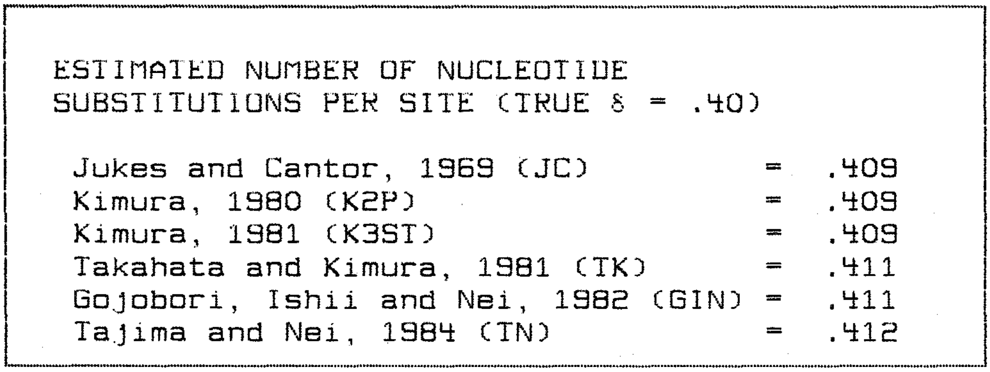
B

```
LOCUS         EVOL1           200 bp     DNA                 entered    04-18-88
COMMENT       Simulated evolved sequence
              Total changes     =     40  (Divergence from the original =   .20)
              Multiple changes =     3
              Nucleotide frequencies:
                 Original:   qA = .250    qC = .250    qG = .250    qT = .250
                 Evolved:    qA = .240    qC = .230    qG = .245    qT = .285
BASE COUNT        48 a      46 c      49 g      57 t
ORIGIN        Sequence evolved from ORIG          under JC.SCH   scheme
         1 TaGcaGggag ttcgaaagtt ctAaTAgtgc tcttCatgTc cAcgGtAatg tgcaTCcGaa
        61 Ccggccgaac attatggggc aaaaGtcccC caatTCggcT tagcTtagac cagtaacggt
       121 cggtcttgcT ATGgGtttcg cGtaaTAgct cgTtcctgtA gcacAtAgaa acgtcatTca
       181 ttgtCtttac ATtgtggcga
//
```

C

```
LOCUS         EVOL2           200 bp     DNA                 entered    04-18-88
COMMENT       Simulated evolved sequence
              Total changes     =     40  (Divergence from the original =   .20)
              Multiple changes =     3
               Nucleotide frequencies:
                 Original:   qA = .250    qC = .250    qG = .250    qT = .250
                 Evolved:    qA = .235    qC = .275    qG = .240    qT = .250
BASE COUNT        47 a      55 c      48 g      50 t
ORIGIN        Sequence evolved from ORIG          under JC.SCH   scheme
         1 aaacatCAag ttcgTaagtt Gtgagcgtgc tctttatgac cgTCcAgatg CgcaGacgaa
        61 aTggccgaac CttatCgAgc aaaactcAcC ATTtctggcc tagcataAAA cagtaGcTgt
       121 cggtcttgcc cTgCcCttcg caGGaTcgct cgctccAgtg gcCcttTgaa acgGcaCaca
       181 CtgtgGttTc tatgtggcga
//
```

**Fig. 2.** Example of simulation using the SDSE package. (**A**) Random sequence of 200 nucleotides at equilibrium frequencies (qA = qT = qC = qG = 0.25) generated by the program XSEC to be used as ancestral sequence for the simulation. (**B**) and (**C**) A pair of 'evolved' sequences generated by the program XSIM; each sequence accumulated 40 random nucleotide changes under the conditions imposed by the JC substitution scheme.



```
ESTIMATED NUMBER OF NUCLEOTIDE
SUBSTITUTIONS PER SITE (TRUE δ = .40)

 Jukes and Cantor, 1969 (JC)          =  .409
 Kimura, 1980 (K2P)                   =  .409
 Kimura, 1981 (K3ST)                  =  .409
 Takahata and Kimura, 1981 (TK)       =  .411
 Gojobori, Ishii and Nei, 1982 (GIN)  =  .411
 Tajima and Nei, 1984 (TN)            =  .412
```

**Fig. 3.** Divergence accumulated by the pair of 'evolved' sequences of Figure 2B and C, as estimated by using the program XDIV.

sequence (e.g. random generated).

Option 4 simulates the process of DNA evolution. Following a selected substitution scheme, the program makes the demanded number of random changes on the ancestral sequence, thus generating an 'evolved' sequence, which is written, under a name provided by the user, to the data disk. One may, for example generate two or more 'evolved' sequences (Figure 2B and C) from an ancestral sequence and compare them with the following utility.

Option 5 allows the user to estimate the divergence accumulated by two 'evolved' sequences (Figure 3). Evolutionary distances are estimated through several indexes of evolutionary distance. A matrix of nucleotide pair frequencies can also be directly used as input, thus allowing, for example, the utilization of data from sequences previously aligned by other programs. The results are first shown on the screen, and then written to the file XDIV.RES on the data disk; the user can obtain a hard copy of this file by using the COPY DOS command.

By choosing option 6 one can quit the program and return to DOS.

## Discussion

We have developed a series of programs simulating the evolution of a pair of DNA sequences under any scheme of nucleotide substitution. The simulation is carried out on finite, variable-length DNA sequences through a strict stochastic process, according to the particular substitution rates imposed by each scheme. In previous works the simulation of the nucleotide substitution process was in part deterministic. For example, Gojobori *et al.* (1982) fixed the average rate of nucleotide substitution per site and per unit time as 0.01, and then the number of time units was automatically imposed by this rate value.

All sequence data files used and generated by the SDSE package conform to the standard GenBank database format, thus allowing, on the one hand, the use of any sequence retrieved from this data bank as an ancestral sequence, and, on the other hand, the application of other packages to analyse, manipulate or retrieve simulated sequences. Since all files are written in ASCII, sequences or substitution schemes can be easily handled through standard DOS commands.

## Acknowledgements

This work was financed in part by the CAICYT of the Spanish Government (PR84-1211-CO2-02).

## References

Burks,C., Fickett,J.W., Goad,W.B., Kanehisa,M., Lewitter,F.I., Rindone,W.P., Swindell,C.D., Tung,C.S. and Bilofsky,H.S. (1985) The GenBank nucleic acid sequence database. *CABIOS*, **1**, 225–233.

Gojobori,T., Ishii,K. and Nei,M. (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. Mol. Evol.*, **18**, 414–423.

Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In Munro, H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, 21–123.

Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.

Kimura,M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, **78**, 454–458.

Nei,M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Tajima,F. and Nei,M. (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, **1**, 269–285.

Takahata,N. and Kimura,M. (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, **98**, 641–657.

*Received on August 23, 1988; accepted on November 15, 1988*