Volume 16 Number 5 1988 Nucleic Acids Research

'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers

Christian Marck

Service de Biochimie, Bâtiment 142, Département de Biologie, Centre d'Etudes Nucléaires de Saclay, 91191 Gif-sur-Yvette Cedex, France

Received August 17, 1987; Revised and Accepted November 15, 1987

ABSTRACT

DNA Strider is a new integrated DNA and Protein sequence analysis program written with the C language for the Macintosh Plus, SE and II computers. It has been designed as an easy to learn and use program as well as a fast and efficient tool for the day-to-day sequence analysis work. The program consists of a multi-window sequence editor and of various DNA and Protein analysis functions. The editor may use 4 different types of sequences (DNA, degenerate DNA, RNA and one-letter coded protein) and can handle simultaneously 6 sequences of any type up to 32.5 kB each. Negative numbering of the bases is allowed for DNA sequences. All classical restriction and translation analysis functions are present and can be performed in any order on any open sequence or part of a sequence. The main feature of the program is that the same analysis function can be repeated several times on different sequences, thus generating multiple windows on the screen. Many graphic capabilities have been incorporated such as graphic restriction map, hydrophobicity profile and the CAI plot - codon adaptation index according to Sharp and Li (1). The restriction sites search uses a newly designed fast hexamer look-ahead algorithm. Typical runtime for the search of all sites with a library of 130 restriction endonucleases is 1 second per 10000 bases. The circular graphic restriction map of the pBR322 plasmid can be therefore computed from its sequence and displayed on the Macintosh Plus screen within 2 seconds and its multiline restriction map obtained in a scrolling window within 5 seconds.(*)

For the past few years, the now familiar graphic user's interface of the Macintosh (2) has established itself as the most friendly way to communicate with the machine. On the other hand, the availability of powerful C compilers has now raised the Macintosh among the most sophisticated microcomputers to program. The DNA Strider program is therefore the result of an effort to take advantage of both the friendly interface and power of this machine for the daily handling of DNA and protein sequences. An overview of the program organisation and philosophy is followed by a short description of some menus and original commands. Finally, the newly designed algorithm for a fast search of the restriction sites is exposed.

DNA Strider is a window-based application that makes full use of the Macintosh style editing facilities. The overall operation of the program is fairly simple: when the user opens a sequence, it is displayed in a Sequence Worksheet window (fig. 1) where it can be further edited and analyzed. Editing of the sequence is then performed by the "File", "Edit" and "Find" menus; a fourth menu, "Conv" (Convert) allows to convert sequence from one type to another one. As a matter of fact, DNA Strider can handle four types of sequences: 1/ DNA sequences in which one

(*) DNA Strider is available (with a detailed user's guide) from the author as a Macintosh application at no charge. Please, send a formatted 800K disk and a self-addressed mailing label.

© IRL Press Limited, Oxford, England.

can use the "*" character as an undetermined base. 2/ DNA Consensus sequences that use the complete degenerate DNA alphabet (A,B,C,D,G,H,K,M,N,R,S,T,V,W and Y, see ref. 3). 3/ RNA sequences. 4/ Protein sequences in one-letter code. Up to 6 different sequences (of any type) can be opened, edited and analyzed simultaneously. Analysis of a sequence, or of a selected part of a sequence, is performed by the commands of the "Enz." (Enzymes) and "A.A." (Amino Acids) menus. The various commands of these two menus generate new windows as detailed later. A short description of the different menus as well as some window and runtime examples are given below (all runtime data given in this paper apply to the Macintosh Plus).

The "File" menu includes all standard Macintosh file commands plus the following: "Duplicate" (gives an unnamed copy of a sequence), "Save a Copy", "Auto Save" (the sequence is automatically saved every two minutes) and "Revert to Saved" (all changes made since the last "Open" command are canceled). Two more commands useful for sequences housekeeping on disks have been also implemented: "Delete" and "New Folder".

The "Edit" menu is used to edit both the sequence itself and the comment field. Editing through different sequences is of course achieved with the standard Cut/Copy/Paste/Clear commands. A "Keep" command is also present. Unlike "Clear", this command clears the parts of the sequence that are not selected.

The "Find" menu is exclusively devoted to sequence editing and includes various jump and search commands. Among them the "Find ORF" command turns out to be very useful: this command moves the selection to the successive open reading frames encountered. Coding sequences can be therefore easily localized and selected for subsequent translation analysis. Three switches allow the various possible searches to be performed either downwards or upwards, as case sensitive or unsensitive and also as phase sensitive or unsensitive.

The "Conv" menu allows to convert a sequence into another one, say DNA into RNA, RNA into protein. Proteins translated from both DNA strands can be directly obtained using the "Protein 5'->3'" or "Protein 3'<-5'" commands. It should be noticed that the conversion is not performed "in place". Instead, a new Sequence Worksheet window is created so that the original sequence is still available. Additional commands are applicable to DNA sequences: "Circularize" is a toggled command that changes the DNA sequence topology whereas "Origin" allows to renumber the sequence (with negative base numbers if desired).

DNA Strider makes extensive use of the Macintosh multi-window graphic interface and a few words are necessary to explain how the commands of the analysis menus "Enz." and "A.A." generate windows. Each of these commands generates a different type of window, e.g. the "KD Hydrophobicity" command creates a window that displays the hydrophobicity profile of the target protein sequence according to Kyte and Doolittle (4). However, several hydrophobicity profiles of different proteins can be obtained together, each in a different window, by running the command several times. All analysis commands behave similarly, generating a new window every time they are invoked and this powerful feature offers to the user many comparison possibilities. Three examples follow: First, two (or more) representations of a same sequence can be displayed at the same time and therefore simultaneously examined, e.g. a multiline restriction map, a list of absent sites and a translation. Second, the user can ask for, say a 3-phase translation of its current DNA sequence, modify this sequence, ask again for a 3-phase

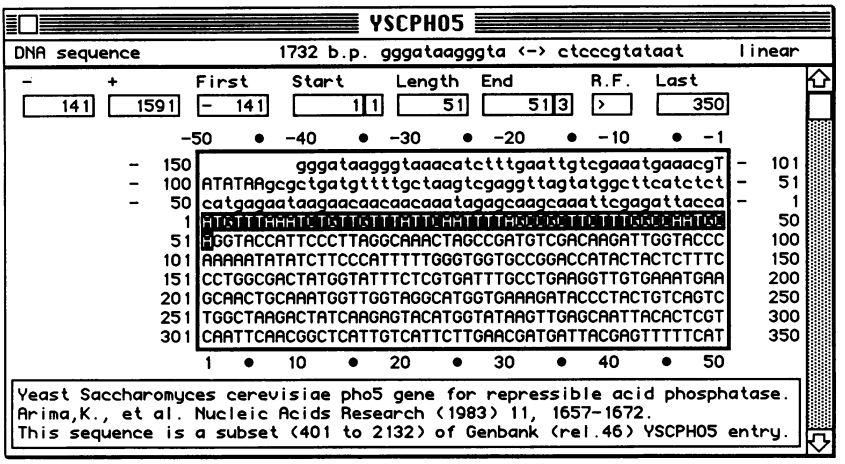

Fig. 1 Example of a Sequence Worksheet window. The line below the title bar displays the sequence type, its full length, the 12 first and 12 last bases and its topology. Immediately above, a row of ten "counters" display various data about the sequence: "-": the number of "negative" bases (i.e. the bases on the 5' side of the origin), "+": the number of "positive" bases, "First": the rank of the first base actually displayed, "Start" the rank of the first selected base with its phase, "Length": the number of selected bases, "End": the rank of last selected base also with its phase, "R.F.": the reading frame status of the selection and "Last": the rank of the last base displayed. If no selection holds, the "Start", "Length", "End" and "R.F." counters are replaced by a single "Position" counter that gives the insertion point position and phase. All these counters are updated whenever needed and therefore while the user types in the sequence or expands the selection with the mouse. The two lateral scales also scroll while the user scrolls the sequence and the two top and botton scales are updated to match accordingly. The wide rectangle at bottom is a 3-line editable comment field which is saved with the sequence. The sequence displayed here encompasses the bases 401 to 2132 from the GENBANK entry YSCPHO5. The origin has been purposely moved to the start of the PHO5 coding sequence. High case letters have been chosen to emphasize the PHO5 coding sequence as well as the TATA box at -101. The fraction of the sequence coding for the signal peptide has been selected with the mouse and the start of an open reading frame is indicated accordingly in the "R.F." counter.

translation and then compare the results of the two translations since both translation windows are still present on the screen. Third, one can convert the current DNA sequence into a protein sequence and then obtain analysis windows from both sequences.

The commands of the "Enz." menu perform various restriction analyses including multiline restriction map and graphic restriction map as shown in fig. 2. Some translation analyses found in the "A.A." menu are given as examples in fig. 3 and 4. The codon adaptation index plot (CAI) allows one to distinguish which parts of a coding sequence use rare codons (1). One can notice in fig. 4 that the 26-fold heptamer repeat at the -COOH end of the B220 subunit of the yeast *S.cerevisiae* RNA polymerase B (6) correlates with a use of rare codons. Scrolling windows for 1-, 3- and 6-phase translation are also available. Finally, besides Sequence Worksheet and analysis windows, DNA Strider also offers to the user some reminder windows, an example of which is given in fig. 5.

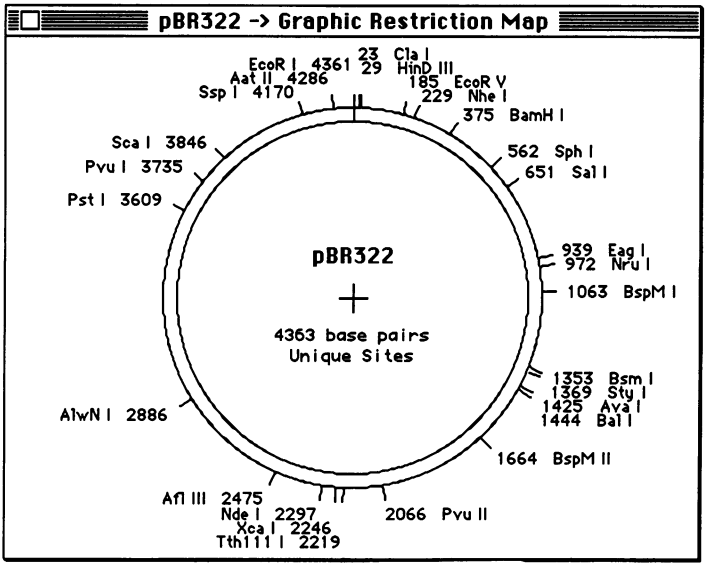

Fig. 2 Example of Graphic Restriction Map window. The restriction enzyme library includes 130 enzymes all with a different recognition sequence. The search for all the sites (using the hexamer look-ahead algorithm), the geometrical computation of the map and its display are all performed in 2 seconds for the pBR322 sequence - 4363 bases (pBR322 entry in GENBANK rel. 46). The limiting step is the use of trigonometric functions since the computation time does not increase linearly as a function of the sequence length: The map of the murine mitochondrial genome (16295 bases) is obtained in 2.8 seconds. Optional dialog allows one to include more sites in the map. For linear sequence, a linear map is produced.

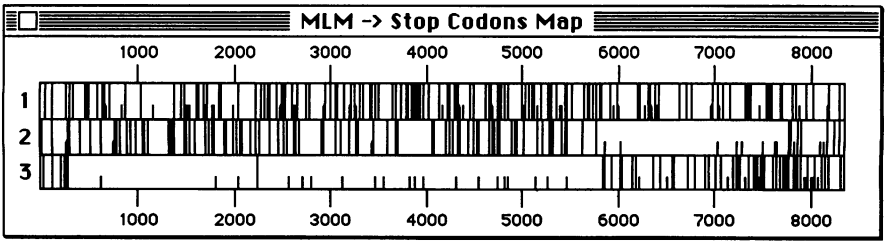

Fig. 3 Example of a Stop Codons Map window. The three direct reading frames are represented as usually with stop codons (TAA, TAG or TGA) as full bars and ATG as interrupted bars. The sequence represented here is that of the Moloney murine leukaemia virus - 8332 bases (MLM entry in GENBANK rel. 46) which displays the three large reading frames of the gag, pol and env polyproteins. A compression factor of 6 is used so that the whole sequence fits into the 512-pixel wide screen of the Macintosh Plus/SE. This window that sums up the status of the three reading frames over 8 kB, is obtained in 3 seconds. An analogous 6-phase display is also available.

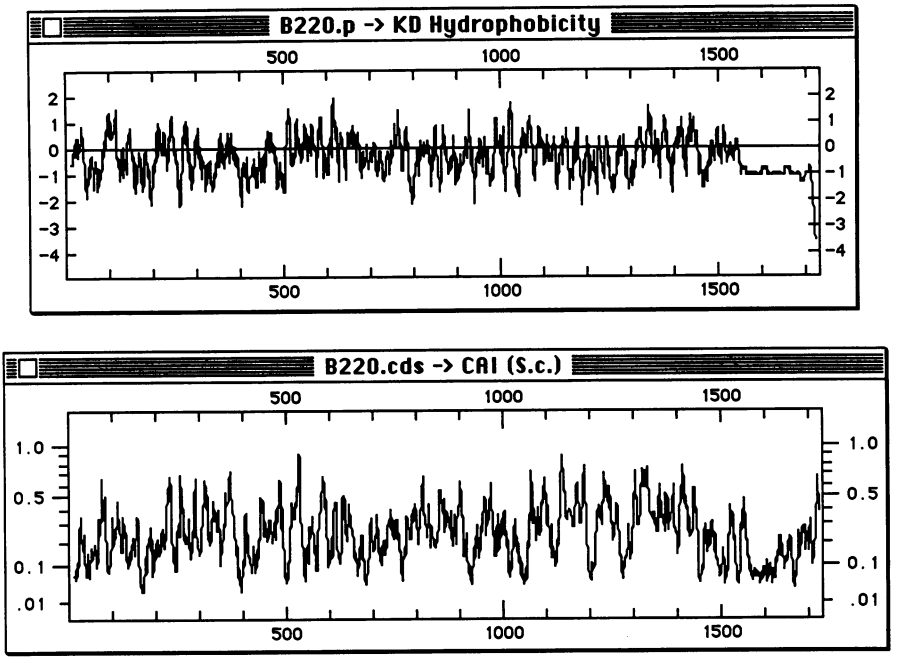

Fig. 4 Examples of a Kyte-Doolittle Hydrophobicity plot window (top) and Codon Adaptation Index plot (bottom). These two representations are applied here to the same sequence: the largest subunit (B220) of the RNA polymerase B (II) of Yeast *Saccharomyces cerevisiae* - 1726 amino acids, 5178 bases - YSCRPO21 entry of GENBANK rel. 46). This protein exhibits at its carboxylic end an unusual "tail" made of a 26-fold heptamer repeat (amino acids 1532 to 1713). While this repeat displays a monotonous hydrophobicity value (top), one can see that the protein uses rare codons around position 1600 as shown by the low CAI value (bottom). The hydrophobicity is computed from the protein sequence according to the coefficients proposed by Kyte and Doolittle (4). The CAI plot (Codon Adaptation Index) is computed from the DNA sequence (since the codons have to be known) using the data proposed by Sharp and Li for *Saccharomyces cerevisiae* proteins (1). A quadratic scale is used to expand the plot in the low CAI values. For both plots, an averaging window of 11 amino acids and a compression factor of 4 are used here. This hydrophobicity plot is obtained in 4 seconds and the CAI plot also needs 4 seconds. An hydrophilicity plot according to Hopp and Woods (5) as well as a CAI plot adapted to *E. coli* proteins (1) are also available.

Hexamer look-ahead restriction site search : A tetramer look-ahead algorithm has already been presented (7) that uses the following observation: among the known restriction sites, around half of the possible tetramer configurations never appear at the 5' end of restriction site sequence. This requires one to build up a look-ahead table of 256 (4^4) entries that tells, for every possible tetramer encountered in the sequence, whether the search can be skipped or not. Taking advantage of the large amount of core memory available in the Macintosh Plus, we have devised a more efficient hexamer look-ahead algorithm. A 4096 (4^6)-entry table is built, the null elements of which indicate, as in the tetramer look-ahead search, that no site starts at the position where

Amino Acids Data

name	code	side group	MW	HW	KD
isoleucine	I ile	-CH(CH3)-CH2-CH3	113.1	-1.8	4.5
valine	V val	-CH-(CH3)2	99.1	-1.5	4.2
leucine	L leu	-CH2-CH(CH3)2	113.1	-1.8	3.8
phenylalanine	F phe	-CH2-phi	147.1	-2.5	2.8
cysteine	C cys	-CH2-SH	103.0	-1.0	2.5
methionine	M met	-CH2-CH2-S-CH3	131.0	-1.3	1.9
alanine	A ala	-CH3	71.0	-0.5	1.8
unknown	X ---		0.0	0.0	0.0
glycine	G gly	-H	57.0	0.0	-0.4
threonine	T thr	-CH(CH3)-OH	101.0	-0.4	-0.7
serine	S ser	-CH2-OH	87.0	0.3	-0.8
tryptophan	W trp	-CH2-indole	186.1	-3.4	-0.9
tyrosine	Y tyr	-CH2-phi-OH	163.1	-2.3	-1.3
proline	P pro	[N]-(CH2)3-[CH]	97.1	0.0	-1.6
histidine	H his	-CH2-imidazole	137.1	-0.5	-3.2
asparagine	N asn	-CH2-CONH2	114.0	0.2	-3.5
aspartic acid	D asp	-CH2-COOH	115.0	3.0	-3.5
glutamic acid	E glu	-CH2-CH2-COOH	129.0	3.0	-3.5
glutamine	Q gln	-CH2-CH2-CONH2	128.1	0.2	-3.5
lysine	K lys	-(CH2)4-NH2	128.1	3.0	-3.9
arginine	R arg	-(CH2)3-NH-CNH-NH2	156.1	3.0	-4.5

Fig. 5 This reminder window sums up some useful data and values concerning amino acids. The user can re-order the lines of this array according to the content of any column (but the "side group" one) just with a mouse click into the desired column. Here, the window has been re-ordered according to the Kyte and Doolittle hydropathy values (last column). Other reminder windows concerning the DNA degenerate alphabet, the genetic code, and the restriction sites library are also available.

the corresponding hexamer is found in the sequence. Furthermore, non null elements of the same table give access to as many lists of a few restriction sites that need to be examined. For example, whenever the CAGCTG hexamer is met, the associated list contains only Pvu II (CAG'CTG), NspB II (CMG'CKG) and AlwN I (CAGNNN'CTG). Since the whole library does not need to be scanned, this results in better efficiency. During the making of the hexamer look-ahead table, a problem might arise from the fact that not all restriction sites are exactly 6 bases long, but this can be overcome:

1/ Sites shorter than 6 bases are continued up to 6 bases with N's (N is any base i.e. A or C or G or T) while bases beyond the 6th one are provisionally disregarded for longer sites. All hexamer combinations are then generated by replacing N's and other degenerate bases found in the consensus sites by their successive equivalent (note that a site as A'CGT (Mae II) generates 16 different hexamers, a site as CACNNN'GTG (Dra III) generates 64). This process generates around 1200 hexamer/enzyme sets for a library of 130 enzymes.

2/ These sets are then sorted for increasing hexamer values with the Quicksort algorithm (8) (the code used to compute the hexamer values is A=0, C=1, G=2 and T=3, therefore AAAAAA has 0 for value and TTTTTT 4095).

3/ Finally, one builds the 4096 entry array, the non null elements of which point to as many lists of restriction sites to look for when a given hexamer is found in the sequence. For the same 130-enzyme library, it appears that 3230 hexamer configurations are never found at the beginning of a site.

Various improvements (not detailed here) allow this algorithm to support non-palindromic as well as palindromic sites and also the presence of empty bases (noted as *) in the searched sequences. The above operations would need to be performed only when the restriction site library is updated; however, since it needs only 4 seconds, the whole process is performed every time the program starts up. This allows one to store the restriction sites as a separate TEXT file that can be easily updated with any word processing program.

It should be noted that this process makes a particular use of the bank segmentation principle: the usual situation is reversed since, instead of looking into the DNA sequence for the sequence of a given restriction site, we look whether a given hexamer (from the DNA sequence) is found in the "bank" formed by the restriction site library. The whole process exposed here may seem complex but the pay-off is great: the entire pBR322 sequence is searched for 130 restriction sites within 0.45 second.

DNA Strider has been designed as a fast and easy to use tool. It is hoped that the runtime and display examples given above show that these two goals were reached. The key feature of the program is that several analysis windows of any sequence, or of a selected part of a sequence, can be obtained simultaneously, moved on the screen and compared. The size of the screen used is therefore a limiting factor and newly available 1024x980 screens are of course best suited for DNA Strider. The program, in its present state is not an end and will be completed by other analysis possibilities such as sequence searches and comparisons, dot matrix analysis and alignments.

System and Methods DNA Strider has been developed with the Macintosh Programmer's Workshop (MPW) and the Macintosh Workshop C Compiler which are new professionnal development tools from Apple. This C compiler supports the Berkeley 4.2 BSD VAX implementation of the PCC (Portable C Compiler). The program code occcupies around 120 K. It requests the 128 K (or 256 K) ROM and Finder 5.5/System 4.1 (or Finder 6.0/System 4.2 to run with MultiFinder). DNA Strider has been tested with the Radius FDP (9) and Megascreen (10) displays. Other third-party screens should also work properly. Both the ImageWriter I (8 or 15"), II and LQ as well as the LaserWriter Plus are supported. DNA Strider uses an original binary format to store the sequences but has also capability to read and write sequences in ASCII TEXT format compatible with word processing and communication programs. Sequences can be passed to and from other applications through the Clipboard.

ACKNOWLEDGEMENTS I thank A. Sentenac and P. Fromageot for interest in this work. I wish also to thank P. Jahu (Apple Computer France) for permission to use and for advice in handling the MPW and the Macintosh Workshop C Compiler.

REFERENCES

1- Sharp, P.M and Li, W-H. Nucl. Acids Res. (1987) **15**, 1281-1295.
2- Apple, ImageWriter and LaserWriter are registered trademarks of Apple Computer,Inc. Macintosh, Finder and MultiFinder are trademarks of Apple Computer, Inc.
3- Cornish-Bowden, A. (1985) Nucl. Acids Res. **13**, 3021-3030.
4- Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. **157**, 105-132.

5- Hopp, T.P. and Woods, K.R. (1981) Proc. Nat. Acad. Sci. **78**, 3824-3828.
6- Allison, I., Moyle, M., Shales, M. and Ingles, C.J. (1985). Cell **42**, 599-610.
7- Marck, Ch. (1986) The Applications of Computers to Research on Nucleic Acids III, IRL Press, Oxford, 583-590.
8- Hoare, C. (1962) Computer J. **5**, 10-15.
9- Radius FDP is a registered trademark of Radius Inc.
10- Megascreen is a registered trademark of Mega Graphics ICS.