SEQUIN

A Program to Assign Nucleic Acid SEQUences INteractively

Nadrian C. Seeman

Department of Chemistry
New York University
New York, NY 10003 USA

PHONE: 212-998-8395
FAX: 212-260-7905
EMAIL: Ned.Seeman@NYU.edu

1990
(Updated to include SPECIAL command, 1993)

This program is a new version of an older sequence-symmetry minimization algorithm that designs sequences so as to maximize control over the secondary structure of branched nucleic acids. The program works by keeping track of the vocabulary elements (sometimes called CRITON's--Seeman, N.C., J. Theor. Biol. 99, 237-247 (1982)) used for the sequence. The program is menu-driven, but the following instructions will probably help the user get off the ground with an introduction to the terminology and the commands (entered in all-CAPS) that are used.

The logic behind this program is described in N.C. Seeman, The Journal of Biomolecular Structure and Dynamics 8, 573-581 (1990).

DISCLAIMER: As far as I know, the program is Bug-free, but of course this is a pretty fantasy. I would appreciate any bugs you discover being brought to my attention.

Structure Assignment.

Nucleic acid structures based on branched or linear DNA molecules are assigned by defining a set of arms and then defining the connectivity between them. Each arm has two strands, 1 and 2. The 3' end of strand 1 of any arm is at the junction. It will be linked to the 5' end of strand 2 of an adjacent arm, to make a continuous strand that forms one corner of the junction. The commands associated with structure assignment are sequence independent.

Caveat: The program is not smart enough to deal with non-integral turns of DNA between double crossovers. Eschew these.

Relevant commands:

NEWARM Define a new arm of a given length. If only a single arm is defined, it may be 15000 nucleotide pairs long. If more arms are defined, the program assumes a branched system is involved, and 150 arms, each 100

nucleotide pairs long, may be defined. Do not mix single and branched molecules with the ADD command.

LINK3 Establish a junction linkage (bonding) between two arms, 3' end of strand 1 (1st arm) to 5' end of strand 2 (2nd arm).

LINK5 Establish a ligation linkage between the two ends of two arms distal from their junctions. Thus, the 5' end of strand 1 of the first arm is bonded to the 3' end of strand 2 of the second arm, and the 3' end of strand 2 of the first arm is bonded to the 5' end of strand 1 of the second arm.

SINGLE Make a given arm a single strand. This allows for more intricate structures, involving loops, hairpins, single-stranded knots, etc. A flag is set .TRUE. when the arm is a single-strand. The complements to single-stranded regions are treated as being in use. This command toggles.

LOOPS Establish the connections between single strands and the rest of the structure.

ADD5 Add positions to the 5' end of an arm. Positions are measured from the 5' end of strand 1.

ADD3 Add positions to the 3' end of an arm.

INSERT Insert positions within an arm.

DELETE Delete positions within an arm.

ZAPARM Delete an arm.

ZAPAC Zap the active molecule--equivalent to starting the program over.

STRANDS Convert the arm structure to strand structure. This command finds all the connections between strands designated by the LINK3, LINK5 and LOOPS commands and outputs the strand sequences to the STRANDOUt unit (see housekeeping commands). This is where you will get into trouble if you do not eschew non-integral helices between double crossovers.

Sequence Assignment.

Sequences are assigned to strand 1 of an arm. Strand 2 is assigned complementarily. Usage of vocabulary elements (CRITONs) is calculated automatically after manual and some automatic sequence assignments.

Relevant Commands:

SET Set the sequence of particular residues. A, T, G, C or N (unspecified) may be assigned for DNA, and A, U, G, C or

N may be assigned for RNA.

DNA Assign DNA residues--the default option.

RNA Assign RNA residues.

AUTO While SETting the sequence, display the usage of the 5' vocabulary elements (CRITONs) that are your choice. The ones shown will be all elements of length MNSHOW through MXSHOW. Thus, if MNSHOW = MXSHOW = 3, and you are setting the 4th nucleotide of a strand whose first 3 are ATC, the previous usage of TCG, TCA, TCT and TCC will be displayed. Palindromic sequences are listed as 1000 + the previous usage, as a warning. This command functions as a toggle.

MXSHOW Maximum size vocabulary elements (CRITONs) to display with the AUTO option with the SET command.

MNSHOW Minimum size vocabulary elements (CRITONs) to display with the AUTO option with the SET command.

CRUNCH Automatically set the sequence of a short (?) stretch of DNA or RNA. Use CRITON after assignment by CRUNCH, just to be safe. Plan to use this command a lot.

SPECIAL Read (only once) a special file, RSITES.DAT, to specify the maximum number of times that particular sequences can appear. The first line of the file contains the number of lines (up to 100) to follow. Each succeeding line contains the maximum number of occurrences of the sequence, the length of the sequence, and the sequence, in (2I4,2X,50A1) format. This capability makes sure that a particular site (e.g., restriction site) doesn't show up inadvertently.

CRITON Calculate the usage of vocabulary elements.

PERMUTE Permute the sequence through one of the eight transformations (including identity) that preserve sequence-symmetry.

AUTOJUNK Display JUNKONs when SETting the sequence. A JUNKON is a set of base pairs that flank a junction, whose sequence one wants to monitor to avoid putting in a sequence that can undergo the branch migration isomerization reaction. The JUNKON can contain several levels--levels are just steps of removal from the junction; the 1st level corresponds to the ultimate bases actually flanking the junction, the 2nd level corresponds to the penultimate bases, the 3rd level to the antepenultimate bases, etc. This command functions as a toggle.

JUNKON Define JUNKONS that will be displayed during the SET command when the AUTOJUNK command is executed. Only enter the ultimate bases.

SHOJUNK Display the JUNKONS that have been defined.

ZAPJUNK Eliminate a JUNKON that has been defined.

File Structure.

The program contains a file system that permits 32 structures to be stored. If you need more files, make more permanent files, renaming them as you need them. Be careful as you get near the limit, because you might get the message that the directory is full, and you cannot save the active structure. The external file is named 'MOLDES.DAT'. A dummy file (containing only junction J1) that can be expanded is provided with the program. When sent, it is named MOLDUM.DAT, and should be renamed immediately.

Relevant Commands:

SAVE Save a structure onto the disk. Do this a lot, regardless of your system reliability.

ERASE Erase a structure from the disk.

RECALL Recall a structure from the disk. Warning: This will zap the active molecule.

ADD Add a structure from the disk to the current active molecule, to make a bigger molecule.

DIR List the directory of molecules on disk.

The program will also seek the file RSITES.DAT, associated with the variable SPECFL (initialized to 9) for the SPECIAL command. An arbitrary file will be supplied with the program, which can be modified for your application.

Molecular Analysis.

The program contains a number of commands to allow the user to see what he/she or the program has done. These are important in figuring out whether a sequence is a good one to synthesize or not.

Relevant Commands:

SHOW Display the sequence of the structure.

SHOARMS Select a numerically contiguous subset of the entire structure whose sequence is to be displayed.

SHOALL Show the sequence of the entire structure, cancelling out the SHOARMS command.

UCR Display vocabulary element (CRITON) usage.

STRING Find the place a given sequence is in the molecule. The location given is the 5' base of the sequence which may

wind up being the 0'th or negative position on a given arm.

SHOGC Display GC content of arms and total structure.

INVEREPS Seek inverted repeats in strands.

MISMATCH Seek all pairings between all strands.

TEMP Select the temperature for energy calculations.

NRGARM Estimate the energy of a given arm. DNA or RNA parameters will be used, depending on the last DNA or RNA command given.

NRGSEQ Estimate the energy of a given sequence. DNA or RNA parameters will be used, depending on the last DNA or RNA command given.

Housekeeping.

There are a few commands that are just used to control input and output and the program.

Relevant Commands:

EXIT Exit the program.

HELP Display the available commands.

NEWIN Change the input unit (defaults to 5).

NEWOUT Change the output unit (defaults to 6).

ECHO Echo commands.

STRANDOUt Select the unit to which strands are output (defaults to 4).

THAT'S IT. I THINK EVERYTHING ELSE IS SELF-EXPLANATORY, OR IS EXPLAINED WHEN THE COMMANDS ARE EXECUTED. IF THERE IS SOMETHING YOU DO NOT UNDERSTAND, PLEASE GIVE ME A CALL.

GOOD LUCK!