Molecular BioComputing Suite: A Word Processor Add-In for the Analysis and Manipulation of Nucleic Acid and Protein Sequence Data

BioTechniques 31:1306-1313 (December 2001)

Patrick Y. Muller, Etienne Studer, and André R. Miserez
University of Basel, Basel,
Switzerland

ABSTRACT

In all fields of molecular biology, researchers are increasingly challenged by experiments planned and evaluated on the basis of nucleic acid and protein sequence data generally retrieved from public databases. Despite the wide spectrum of available Web-based software tools for sequence analysis, the routine use of these tools has disadvantages, particularly because of the elaborate and heterogeneous ways of data input, output, and storage. Here we present a Visual Basic-encoded Microsoft® Word® Add-In, the Molecular BioComputing Suite (MBCS), available at the BioTechniques Software Library (www.BioTechniques.com). The MBCS software aims to manage and expedite a wide range of sequence analyses and manipulations using an integrated text editor environment including menu-guided commands. Its independence of sequence formats enables MBCS to be used as a pivotal application between other software tools for sequence analysis, manipulation, annotation, and editing.

INTRODUCTION

As a consequence of the rapid accumulation of sequence information in all fields of molecular biology and biomedicine, researchers increasingly depend on computers to retrieve, analyze, and edit nucleotide and protein sequences electronically. The huge amount of publicly available data sets of complete genomes, the requirement for the routine use of recombinant DNA technology, and the increasing efforts in sequencing of disease genes make it necessary to store and handle nucleotide and protein sequences electronically. Apart from all kinds of sequence alignments (10,21), common tasks in computational molecular biology include the identification of DNA and protein subsequences with specific functions such as cleavage sites of restriction endonucleases (22), consensus binding sites of transcription factors (13), or target sites for posttranslational modifications. As part of such work, DNA sequences frequently need to be reformatted [e.g., in standardized formats such as FASTA (20)], numbered according to the nucleotide positions, or transformed to their complementary, reverse, or antisense counterparts. In addition, DNA and protein sequences often need to be interconverted to each other (i.e., translated or reverse-translated). On the protein level, common physicochemical properties such as molecular weights, isoelectric points (IP), molar absorption coefficients, compositions of amino

acids, and their molar percentages are often required. Furthermore, graphical plots of window-averaged biochemical, immunological, and genetic parameters as a function of the individual amino acid residues of a protein have become established procedures in molecular biology (14,29).

Nucleotide and protein sequence analysis and manipulation can be performed using a variety of software tools. These tools include console-based applications (28), Web-based interfaces to server-side PERL or C++ scripts (33), or sophisticated software suites running on UNIX®-like operating systems (8). Although console and Web-based software tools have been introduced increasingly, the routine use of data input, output, and storage in these programs may be inconvenient. Furthermore, performing a series of analyses with different software tools, the sequence data to be analyzed is required in a specific format and frequently needs to be reformatted when proceeding from one analysis tool to another or when the sequence data have to be stored. Although there are several, even Web-based, reformatting tools like, for example, *ReadSeq* (<http://bimas.dcrn.nih.gov/molbio/readseq/>), switching between different sequence formats usually is cumbersome. In addition, sophisticated software suites providing an integrated environment are often expensive and, thus, beyond the means of small research groups.

Within the last years, several com-

BioComputing/BioInformatics >>>>

puter applications aiming to provide a convenient way to perform common procedures used in computational molecular biology have been developed (2,3,9,12,17,23–26). Here we present a software application, the *Molecular BioComputing Suite (MBCS)* that was implemented in Microsoft® Visual Basic for Applications. *MBCS* operates as an add-in for Microsoft Word® and solves the limitations addressed, including the financial considerations. *MBCS* is an improvement over already existing tools of its kind, as a huge range of operations can be conveniently performed within Microsoft Word with minimal requirements of time. The main objective is to expedite and simplify the analysis and manipulation of nucleic acid and protein sequence data in an integrated text editor environment including menu-guided commands.

MATERIALS AND METHODS

The *MBCS* source code is entirely coded in the object-oriented programming language Visual Basic for Applications and is contained in one module and nine graphical user forms, all integrated in a Microsoft Word add-in file. The *MBCS* software itself comprises a total of 29 individual program components. Their compilation is automatically accomplished with the Visual Basic compiler upon running one of the *MBCS* programs. The Visual Basic compiler is an integrated constituent of Microsoft Visual Basic for Applications and, thus, available on any system where one of the Microsoft Office Applications is installed. The *MBCS* add-in runs within Microsoft Word (version 97 or later for Microsoft Windows® and version 98 or later for Apple® Macintosh®) and, thus, can be operated on any system where this widely used word processor is installed. All common features of the word processor environment can be used for editing, annotation, and storage of sequence data. All programs of the *MBCS* software can be directly executed within Microsoft Word, and the graphical output is generated by a special Microsoft Excel® file (plot.xls). To have permanent access to these programs, it is recommended to copy the *MBCS* application file into the Microsoft Word startup directory (Menu *Tools, Preferences, File Locations*) or to temporarily activate the file as an add-in (Menu *Tools, Templates and Add-Ins, Global templates and add-ins*). All programs can then be launched from an additional menu generated by *MBCS* (Figure 1). This menu is added to the Microsoft Word menu bar. Furthermore,

tosh®) and, thus, can be operated on any system where this widely used word processor is installed. All common features of the word processor environment can be used for editing, annotation, and storage of sequence data. All programs of the *MBCS* software can be directly executed within Microsoft Word, and the graphical output is generated by a special Microsoft Excel® file (plot.xls). To have permanent access to these programs, it is recommended to copy the *MBCS* application file into the Microsoft Word startup directory (Menu *Tools, Preferences, File Locations*) or to temporarily activate the file as an add-in (Menu *Tools, Templates and Add-Ins, Global templates and add-ins*). All programs can then be launched from an additional menu generated by *MBCS* (Figure 1). This menu is added to the Microsoft Word menu bar. Furthermore,

Figure 1. Screen shot of the *MBCS* toolbar and the *MBCS* output form. The floatable toolbar with the *MBCS* menu, subdivided into three sections, allows a rapid and menu-guided launch of all programs of the software. Screen shot of the output form containing output data of the program *Translation of DNA Sequence*.

BioComputing/BioInformatics >>>>

to provide rapid access, *MBCS* generates a floatable menu bar containing buttons for the launch of the most frequently used programs (Figure 1). To be able to run *MBCS* in Microsoft Word 2000 and, in the meantime, to maintain macro-virus protection for unknown documents, it is recommended to activate the "Trust all installed add-ins and templates" checkbox (Tools/Macro/Security/Trusted Sources).

For evaluation and validation purposes, the *MBCS* software was subjected to two beta testing phases in one molecular genetics and two unrelated molecular biology laboratories. The *MBCS* file referred to in this article can be downloaded free of charge from the Software Library on the *BioTechniques* Web site (www.BioTechniques.com). Moreover, the unencrypted Visual Basic source code of the module and the graphical user forms can be freely edited and extended to develop further applications.

RESULTS AND DISCUSSION

Structure, Features, and Characteristics of the *MBCS* Software

The individual program components of *MBCS* are divided into three sections. The first section includes utilities for import, manipulation, and formatting of nucleic acid and protein sequences. The second section consists of various programs for the sequence analysis of DNA, RNA, and proteins. The objective of the third section is the calculation of numerous pre- and post-experimental parameters required for common molecular biological methods. The data input requirements for each program of *MBCS* are stated upon launching the respective program.

The data input procedure is basically the same for all programs dealing with the manipulation and analysis of sequence data. First, the sequence stretch to be manipulated or analyzed has to be selected with the mouse pointer; thereafter, the desired program has to be launched, either by selecting an item from the *MBCS* menu or by clicking directly on a button of the menu bar (Figure 1). The sequence stretch selected may contain any char-

acters, digits, or symbols not used in the standard one-letter code for DNA, RNA, or protein sequences. *MBCS* automatically eliminates these characters, digits, and symbols by means of a special program subroutine. Moreover, all programs of *MBCS* are case-insensitive. As a consequence, all kinds of sequence formats are accepted, disregarding whether they contain line breaks, tab stops, spaces, digits, or upper- or lowercase letters. This feature provides highest flexibility and, thus, is useful when working with various sequence formats, including sequences contained in GenBank® files (5) retrieved from the Web. In the case of DNA and RNA sequences, the ambiguous one-letter code according to the International Union of Biochemistry and Molecular Biology (IUB) Commission on Biochemical Nomenclature is supported as well (15).

The output data will be displayed in the output form, a window that opens automatically upon completion of any *MBCS* program (Figure 1). The output form provides the user with the following three options: (i) the output data can be directly printed, (ii) the output data can be copied to the clipboard, and (iii) the output data can be pasted to the end of the active Word document.

Working with *MBCS*

Section 1: tools for sequence formatting and manipulation. Every manipulation and analysis of sequence data requires the desired sequence to be either retrieved from a database or generated by the output of an automatic DNA sequencer. For many applications, GenBank (5) is the sequence database of choice, particularly because of its convenient accessibility on the Web. To handle, manipulate, and analyze a Web-based GenBank file with *MBCS*, the content has to be transferred from the Web browser to Microsoft Word while retaining the layout. This task is managed by the *Import Sequence* program that automatically reformats the sequence in a fixed width font.

For the reformatting of nucleic acid and protein sequences, *MBCS* contains four special programs. Two of them are mainly used to format specific subsec-

tions of a sequence in upper- or lowercase characters, a built-in feature of Microsoft Word. This feature is commonly used to annotate and distinguish intron/exon and translated/untranslated regions. The third program for reformatting is intended to convert any kind of nucleic acid and protein sequence into the standardized and widely used FASTA format (20). Another commonly used layout is the numbering of a sequence according to its nucleotide or amino acid positions, which is accomplished by the program *Numbering of Sequence*.

For the conversion of DNA or RNA sequences to their complementary, reversed, or antisense counterparts, three individual programs, named according to these procedures, are provided.

The program *Stepper* offers the opportunity to mark a subsection of a sequence consisting of a user-defined number of nucleotides or amino acids by automatically moving the cursor forward or backward from the actual position, whereas all non-sequence characters are ignored.

Section 2: tools for the analysis of nucleic acid and protein sequence data. The program *Analysis of DNA Sequence* calculates all kinds of physicochemical parameters of the selected DNA sequence. These parameters include overall length, molecular weight of double and single strands, molar percentages of the individual nucleotides, and various melting temperatures. The melting temperatures of oligonucleotides are determined by four different algorithms: (i) the popular and simple GC-method (27), (ii) the salt-adjusted method of Baldino et al. (4), (iii) the advanced thermodynamic nearest-neighbor method (7), and (iv) the T_p -method that corresponds to the maximum temperature at which efficient PCR amplification is observed (30).

The program *Analysis of Protein Sequence* calculates all kinds of physicochemical parameters of the selected protein sequence. These parameters include overall length, molecular weight, molar percentages of the individual amino acids, IP according to Bjellqvist et al. (6), and molar absorption coefficient in solution at the 280-nm wavelength (19).

To identify specific DNA or protein

subsequences, the *DNA Sequence Finder* and the *Protein Sequence Finder* are provided. They can be used to search for specific DNA or protein sequence stretches, respectively, whereas the ambiguous code for nucleotides and amino acids is supported. In addition to specific amino acids, basic, acidic, nucleophilic, aromatic, and aliphatic side chains can be used as search criteria.

For the inter-conversion of nucleic acid and protein sequences (i.e., translation and reverse translation), the programs *Translation of DNA Sequence* and *Reverse Translation of Protein Sequence* can be used. The output of the program *Reverse Translation of Protein Sequence* generates DNA sequences with synonymous codons by using the ambiguous nucleic acid code.

Established procedures for topological predictions of protein sequences include graphical plots of window-averaged biochemical, immunological, and genetic parameters in function of the individual amino acid residues of a protein. In this context, four programs, named according to the underlying method, can be used for such analyses. The window-averaged hydropathy and the grand average of hydropathy (GRAVY) score are computed according to Kyte and Doolittle (14), the polarity according to Grantham (11), and the antigenicity according to Welling et al. (29). The relative mutability of a specific amino acid corresponds to the reciprocal number of its synonymous codons. The determination of the relative mutability is a useful procedure to predict the effects of potential nucleotide substitution mutations on the protein. Protein regions with a high relative mutability are prone to mutations leading to amino acid substitutions. Thus, such protein regions are valuable candidate regions for mutation screening aiming to identify single-nucleotide polymorphisms (SNPs) (1,18) leading to amino acid substitutions. The GC-content plot of a DNA sequence allows the identification of GC-rich sequences and, hence, the selection of appropriate conditions for the sequencing of those templates on the basis of their GC-content (16). The program *Codon Statistics* calculates the individual frequencies of all of the 64 possible codons of a specific DNA

sequence and the frequency distributions among synonymous codons. This data provides the basis for the recognition of a possible codon bias frequently observed in the coding region of highly expressed genes because of selective pressure towards maximizing the efficiency of translation (31).

Nowadays, in the post-genomic era, many DNA sequencing applications are directed towards a comparison with already existing DNA sequences—an approach frequently used for mutation detection in human molecular and medical genetics. This laborious process is sped up and simplified by the program *Mutation Finder*. This program performs an alignment of two selected DNA sequences by searching for overlapping sequence motifs. Nucleotide substitutions between the two compared DNA sequences (i.e., nucleotides not matching the template) are displayed in the alignment. The *Mutation Finder* is intended to be used for sequences of up to approximately 500 bases, depending on the performance of the user's computer system.

Section 3: benchtop techniques.

This section contains seven programs for benchtop use that all provide a graphical user interface. The objective of these programs is the calculation of various parameters for common molecular biology techniques.

The *Molecular Weight Calculator* determines the molecular weight of any compound on the basis of its chemical formula or atomic composition. The *Concentration Calculator* determines the concentration and further parameters of various kinds of biological samples on the basis of the dilution and the absorbance at specific wavelengths. Thus, this program can be used for the concentration calculation of ssDNA, dsDNA, ssRNA, peptides and proteins, bacterial suspensions, and for the determination of the helix content of proteins based on their circular dichroism measurements (optical ellipticity) (32).

The *Centrifugation Calculator* enables the calculation of the relative centrifugal force, the rotations per minute, or the effective radius of the rotor if any two of these parameters are given.

The calculation of the composition of complex PCR mixtures is accomplished by the *PCR Setup Calculator*.

BioComputing/BioInformatics >>>>

This program calculates all volumes of the individual components that have to be added to the PCR mixture. The input variables for this calculation are the stock concentrations of the individual components and their desired final concentrations in the PCR mixture.

The *Electrophoresis Calculator* is based on the proportional relation between voltage and migration distance and their inverse proportional relation to the migration time when a fragment of given molecular weight is subjected to electrophoresis with otherwise fixed conditions. Applying these relationships, it is possible to calculate voltage, migration distance, or migration time for any electrophoresis once one set of these three parameters has been recorded.

The *Ligation Calculator* enables the calculation of the volumes of the vector and insert stock solutions to be used in the ligation reaction. This is accom-

plished on the basis of the fragment lengths and the molar ratio of fragment to insert as well as the concentrations of the vector and the insert stock solutions.

Using the *Technical Appendix*, the user has the opportunity to look up the human codon table of the genetic code, the genetically encoded amino acids and their residue molecular weights, the IUB one-letter code for ambiguous nucleotides, the Web addresses of important human SNP databases, and finally the references of all methodical and technical publications used to implement *MBCS*.

CONCLUSION

The *MBCS* software described here is a tool to manage, expedite, and simplify common procedures for sequence analysis and manipulation using an in-

tegrated text editor environment including menu-guided commands. The application spectrum of *MBCS* is limited to established procedures using moderate amounts of sequence data and does not intend to replace sophisticated Web- or console-based software suites. Nevertheless, *MBCS* is a useful tool for researchers performing computer applications in the field of molecular biology and biomedicine. Because of its independence of sequence formats, *MBCS* is able to take up a pivotal position between various software tools for sequence analysis, manipulation, annotation, and editing.

ACKNOWLEDGMENTS

The authors thank Harald Janovjak for his expertise and his helpful advice concerning the implementation of

object-oriented program components. Moreover, the authors thank the colleagues who assisted them with their valuable suggestions and discussions, especially those who performed the beta testing of the software. This work was supported by the Swiss National Science Foundation grant no. 3200-049125.96/2. A.R.M. is supported by the Swiss Clinicians Opting for Research (SCORE) A grant no. 3231-048896.96 of the Swiss National Science Foundation.

REFERENCES

- Altshuler, D., V.J. Pollara, C.R. Cowles, W.J. Van Etten, J. Baldwin, L. Linton, and E.S. Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513-516.
- Anbazzhagan, R. and E. Gabrielson. 1999. Spreadsheet-based program for alignment of overlapping DNA sequences. *BioTechniques* 26:1180-1185.
- Anbazzhagan, R., J.G. Herman, K. Enika, and E. Gabrielson. 2001. Spreadsheet-based program for the analysis of DNA methylation. *BioTechniques* 30:110-114.
- Baldino, F., Jr., M.F. Chesselet, and M.E. Lewis. 1989. High-resolution in situ hybridization histochemistry. *Methods Enzymol.* 168:761-777.
- Benson, D.A., M.S. Boguski, D.J. Lipman, and J. Ostell. 1997. GenBank. *Nucleic Acids Res.* 25:1-6.
- Bjellqvist, B., B. Basse, E. Olsen, and J.E. Celis. 1994. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* 15:529-539.
- Breslauer, K.J., R. Frank, H. Blocker, and L.A. Marky. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83:3746-3750.
- Butler, B.A. 1998. Sequence analysis using GCG. *Methods Biochem. Anal.* 39:74-97.
- Cabrera Leon, N. 1999. Word-processor macro for restriction endonuclease analysis. *BioTechniques* 27:1228-1231.
- Gaskell, G.J. 2000. Multiple sequence alignment tools on the Web. *BioTechniques* 29:60-62.
- Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862-864.
- Han, B. and A.H. Tashjian, Jr. 1998. User-friendly and versatile software for analysis of protein hydrophobicity. *BioTechniques* 25:256-263.
- Heinemeyer, T., X. Chen, H. Karas, A.E. Kel, O.V. Kel, I. Liebich, T. Meinhardt, I. Reuter, F. Schacherer, and E. Wingender. 1999. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* 27:318-322.
- Kyte, J. and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105-132.
- Liébecq, C. 1992. *Biochemical Nomenclature and Related Documents*. Portland Press, London.
- Motz, M., S. Paabo, and C. Kilger. 2000. Improved cycle sequencing of GC-rich templates by a combination of nucleotide analogs. *BioTechniques* 29:268-270.
- Muller, P.Y., L. Barella, and A.R. Miserez. 2001. RepGene: a spreadsheet template for the management of reporter gene assays. *BioTechniques* 30:1294-1298.
- O'Dell, S.D., T.R. Gaunt, and I.N. Day. 2000. SNP genotyping by combination of 192-well MADGE, ARMS and computerized gel image analysis. *BioTechniques* 29:500-506.
- Pace, C.N., F. Vajdos, L. Fee, G. Grimsley, and T. Gray. 1995. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* 4:2411-2423.
- Pearson, W.R. 1994. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* 24:307-331.
- Phillips, A., D. Janies, and W. Wheeler. 2000. Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.* 16:317-330.
- Roberts, R.J. and D. Macelis. 1999. REBASE-restriction enzymes and methylases. *Nucleic Acids Res.* 27:312-313.
- Shaw, G. 1995. Protein sequence interpretation using a spreadsheet program. *BioTechniques* 19:978-983.
- Shaw, G. 2000. Useful Microsoft® Word® macros for molecular biologists and protein chemists. *BioTechniques* 28:1198-1201.
- Singal, R. and S.R. Grimes. 2001. Microsoft® Word® macro for analysis of cytosine methylation by the bisulfite deamination reaction. *BioTechniques* 30:116-120.
- Stowe, R.P. and D.L. Pierson. 1996. Spreadsheet macro for setting up PCR assay tubes. *BioTechniques* 20:1088-1089.
- Suggs, S.V., R.B. Wallace, T. Hirose, E.H. Kawashima, and K. Itakura. 1981. Use of synthetic oligonucleotides as hybridization probes: isolation of cloned cDNA sequences for human beta 2-microglobulin. *Proc. Natl. Acad. Sci. USA* 78:6613-6617.
- Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.
- Welling, G.W., W.J. Weijer, R. van der Zee, and S. Welling-Wester. 1985. Prediction of sequential antigenic regions in proteins. *FEBS Lett.* 188:215-218.
- Wu, D.Y., L. Ugozzoli, B.K. Pal, J. Qian, and R.B. Wallace. 1991. The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol.* 10:233-238.
- Xia, X. 1996. Maximizing transcription efficiency causes codon usage bias. *Genetics* 144:1309-1320.
- Yang, J.T., C.S. Wu, and H.M. Martinez. 1986. Calculation of protein conformation from circular dichroism. *Methods Enzymol.* 130:208-269.
- Zhang, J. and T.L. Madden. 1997. PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. *Genome Res.* 7:649-656.

Received 12 June 2001; accepted 10 August 2001.

Address correspondence to:

Patrick Y. Muller
Research Group Cardiovascular Genetics
Department of Clinical-Biological Sciences
University of Basel
Vesalgasse 1
CH-4051 Basel
Switzerland
e-mail: patrick.muller@unibas.ch

For reprints of this or
any other article, contact
Reprints@BioTechniques.com