Computer Program for Automatically Calculating Similarity Indexes from DNA Fingerprints

E. Scott Archer and Frederick C. Leung
The University of Hong Kong, Hong Kong

ABSTRACT

DNA Simdex™ Version 2.1 is a third-generation Windows® computer program that uses sophisticated image-analysis techniques to automatically locate bands on electrophoresis gel images and calculate similarity indexes. Here we give a functional description of the program.

INTRODUCTION

DNA fingerprinting is useful in studies on the relatedness between individuals and species, as it can be used in estimating individual and population homozygosity and relatedness. DNA sampled from an individual is digested by one or more restriction enzymes, producing a collection of DNA fragments of varying lengths, which are resolved using gel electrophoresis. The similarity index between DNA fingerprints, measured as a function of the number of matching vs. non-matching bands, as described by Lynch (2,3) can then be used to make inferences about the genetic structure of populations.

Although helpful in its ability to quickly calculate similarities if band positions were known, the initial version of DNA Simdex™ (5) had no built-in image analysis and thus relied on outside means for determining band locations. Several other image-analysis systems are available that are capable of automated band detection and subsequent calculation of similarity, but the cost of these systems is somewhat prohibitive since they typically require specialized hardware and costly software. In response, we have developed DNA Simdex Version 2.1 (Simdex2.1), a user-friendly graphical analysis program that requires no more expensive hardware than a PC running Windows 95® or NT™ and an image-acquisition device such as a desktop scanner or camera/digitizer.

MATERIALS AND METHODS

A desktop scanner can be used to digitize an autoradiogram or fluorescent photograph, as described previously (1,6,7). Alternatively, a charge-coupled device (CCD) camera-based, image-analysis system can be used, but it must allow saving of digitized images in standard image file formats. For Simdex2.1, the image must be stored as a 256 gray-scale bitmap, in bitmapped (BMP) format, with the wells of the gel located along the top of the image and the lanes running vertically down.

Upon loading the gel image file, two windows are created: one that contains the digitized gel image and one that will show the corrected band positions when they are available. The user is prompted to use the mouse to designate a rectangular region on the gel image, which defines the locations of the wells, the migration front and the outermost lanes. The user can then designate the lanes’ locations, such that the lane-vector runs approximately through the middle of the entire lane.

When the user is finished editing the lane locations, the band-editing dialog pops up. This dialog acts as a sort of control box, from which the user can initiate an automated search for bands, which is performed using a template-matching technique, similar to that described previously (4). In this technique, a small “window” of intensity values along the lane is checked for a match against a relative-value template. The window is iterated along the entire length of the profile, one pixel at a time. For example, if the window size is set at five, call each profile value in the window a, b, c, d and e. A “peak” is detected if the values match the template: \( a \leq b \leq c \geq d \geq e \).

The lane’s background is used as a filter during peak detection on the profile, such that any detected peaks, or local maxima, with values above background are designated as bands. The user can adjust the filter value of the background to compensate for an excessively faint, dark or noisy gel image.

Because the effectiveness of finding bands using template-matching is about 85%, the user can also add, remove and edit band locations manually using the mouse. To aid the user in editing bands, a graph of each lane’s intensity profile can se-
lectively be drawn along the lane. This allows the user to make sure that bands are accurately positioned.

The length of a DNA fragment’s migration through a gel, and consequently its band position, is determined by a combination of factors, including (but not necessarily limited to) its molecular weight and its charge and conformation. Migration length can be influenced by other factors too, such as inconsistencies in the gel matrix and the quantity of DNA loaded into the lane. Variations in migration length can cause mismatches in bands that should otherwise match. While the human observer easily compensates for these distortions, the computer must base its band-matching decisions on discrete pixel locations.

To this end, Simdex2.1 provides a distortion map (Figure 1). It acts as a set of visible contours that run across all of the lanes, representing equal migration length. Using the mouse, the user can modify the location of any contour-lane intersection, which re-maps the migration length of any nearby bands. This turns out to be an effective method of compensating for most gel distortions.

The molecular weight (MW) of a particular band on a gel image is not directly implied by its visible migration position, which in Simdex2.1 is measured in pixels. Simdex2.1 assigns approximate molecular weights to bands using references to lanes that contain a MW standard. The bands of a MW standard are of known molecular weight and thus provide a set of reference points for mapping corrected migration lengths to molecular weights.

First, the user must select the MW standard lanes. After assigning a MW standard from the modifiable MW standard database, the user must assign a molecular weight from the standard to one of the bands in a standard lane. With the user’s assistance, Simdex2.1 maps the rest of the bands in the standard lanes to corresponding molecular weights in the MW standard. Simdex2.1 then calculates the molecular weights of the bands in the rest of the lanes, based on their migration length, as corrected by the distortion map.

Simdex2.1 can calculate the similarity indexes for any set of lanes on a single gel or on different gels. The similarity index is given by Equation 1:

\[ S_{ab} = \frac{2n_{ab}}{n_a + n_b} \]

where \( S_{ab} \) is the similarity index between lanes \( a \) and \( b \), \( n_{ab} \) is the number of bands common to both lanes and \( n_a \) and \( n_b \) are the number of bands in lanes \( a \) and \( b \), respectively. For similarity index calculations on lanes contained in a single gel, the bands’ corrected migration distances are used. However,
when comparing lanes on more than one gel, the molecular weights of the bands involved must be known. This is an obvious requirement, as molecular weight is the only unambiguous quantity available between gels.

Simdex2.1 displays band migration lengths, MWs and any calculated similarity indexes in tabular displays (Figure 2). These data can also be printed or exported to files as American Standard Code for Information Interchange (ASCII) text for import into other software, such as a spreadsheet program. Additionally, work can be saved to a project file at any time for later retrieval. Slated for future releases are the additions of TWAIN scanner support and an open database connectivity (ODBC)-compliant database connection to support similarity calculations across very large data sets.

**AVAILABILITY**

DNA Simdex Version 2.1 is available on the World Wide Web (http://www.qualia.net/genetx/gx_simdex2.1.htm).

**ACKNOWLEDGMENTS**

We thank Dr. Darrell P. Chandler and Mr. Ryan D. Quisenberry for providing DNA fingerprint autoradiograms used in the testing and development of DNA Simdex Version 2.1. This work was partly supported by the US Department of Agriculture (USDA) under Related Service Agreement with the US Department of Energy (DOE) under Contract No. DE-AC06-76RLO 1830 at Pacific Northwest Laboratory (PNL). PNL is operated for DOE by Battelle Memorial Institute. This work was also supported by the Hong Kong Research Grants Council, The University of Hong Kong 196/95M.

**REFERENCES**


Received 9 January 1998; accepted 20 April 1998.

Address correspondence to:
Dr. Frederick C. Leung
Department of Zoology
The University of Hong Kong
Hong Kong
Internet: fcleung@hkucc.hku.hk