NTSYSpc 2.02e Implementation in Molecular Biodata Analysis (Clustering, Screening, and Individual Selection)

Soleiman Jamshidi [1] [+] and Samira Jamshidi[2]

[1] Department of Plant Protection, Miyaneh Branch, Islamic Azad University, Miyaneh, Iran
[2] Islamic Azad University, Miyaneh, Iran

Abstract. NTSYSpc is one of the most popular softwares being used in molecular genetic qualitative data cluster analysis. The present paper is showing how we can integrate this powerful software with Microsoft Office Word and Excel in an innovative method to cluster, screen and more varied individuals selection in a populated group studying. A step by step procedure has been explained here in detail. Using this method may help to select the individuals with the highest possible difference from others in a large population using a logical and mathematical protocol with high accuracy based on the least relativeness with other individuals. The screened individuals can be considered for further assessments such as nucleotide sequencing in the case of biodiversity studying on massive populations.

Keywords: molecular screening, genetic diversity, numerical taxonomy

1. Introduction

NTSYSpc stands for "**N**umerical **T**axonomy **SYS**tem for **p**ersonal **c**omputer" is a system of statistical programs that is used to find and display structure in multivariate data. The program was originally developed for use in biology but has also been widely used in morphometrics, ecology, and in many other disciplines in the natural sciences, engineering, and the humanities (Rohlf, 1998). Perhaps the most common use of NTSYSpc is for performing various types of agglomerative cluster analysis of some type of similarity or dissimilarity matrix (Rolf, 1998 and TariNezhad *et al*., 2005).

Screening is the investigation of a great number of something for instance, people looking for those with a particular problem or feature. Important cases of screening are used in biology and related disciplines such as botany, zoology, genetics, medicine, pharmacology, microbiology, and also in security and etc (Wikipedia, 2011). Screening and subsequently selection of individuals is one of the methods to facilitate intensive study of massive populations. It aims to ignore the most similar individuals and consider the ones with the most difference in biodiversity studies of populations (Tanaka *et al*., 2004). Screening and selection in living things are based on phenotypic and genetic features (Baxevanis and Outellette, 2005). In the past decades, several key advances in molecular genetics have greatly increased the impact of population genetics on biology (Sunnuks, 2000), the methods based on PCR and DNA hybridization is being usied in this filed (Spooner, 2005). The selection criteria are usually established by inaccurate methods with no mathematical, logical bases, depending on the researchers' elegance, preferences, and sometimes pre-judgments. Thus, this inappropriate selection will affect the subsequent results to be less punctual. In huge populations, the initial screening is performed using some methods like RAPD and ISSRs (Wang, 2009).

Softwares such as SPSS and NTSYSpc are the most common softwares being used for clustering based on some methods like UPGMA, COMPELETE, ELEXI, SINGLE, UPGMC, WPGMA, WPGMC and WPGMS (Rolf, 1998). In the software, the individuals are only clustering but there is no way for screening

[+] Corresponding author. Tel.: + 98 (423) 2237040-4; fax: + 98 (423) 2227290.
E-mail address: s.jamshidi@m-iau.ac.ir

and selection. The procedure that is going to be explained is a logic protocol starting from the first step of clustering to accurate selection of the most varied individuals integrating NTSYSpc and Microsoft Office Word and Excel 2010.

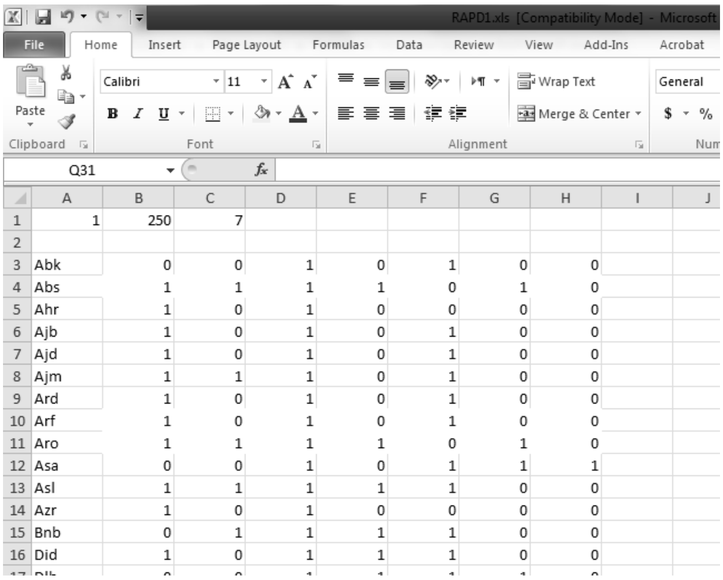
2. Step by Step procedure

2.1. Molecular data preparation and entry

Enter data obtained from on agarose gel as 0 and 1 representing of absence and presence of a DNA band in following order in Microsoft Office Excel 97 ~ 2010:

- Enter each primers data in a separated data sheet in a single file.
- Write always “1” in “A1” cell as a sign of rectangular matrix.
- In “B1” cell write the total number of individuals (isolates, varieties, races, genotypes etc.).
- In “C1” cell write the maximum band number you have got using this primer.
- Leave blank the second raw (A2, B2, C2 …… Z2).
- Write the name, number, code or acronym of individuals starting from “A3” ~ “An”, n = number of individuals.
- Enter 0 and 1 data in “B3” ~ “Z3” for the first individual and “B4” ~ “Z4” for the second one and …
- Save file as arbitrary name using by “Save as” sub-menu in File menu.

> **Important**: if you are using Microsoft Office Excel 2007 and 2010, save file as “Excel 97-3003 document”.

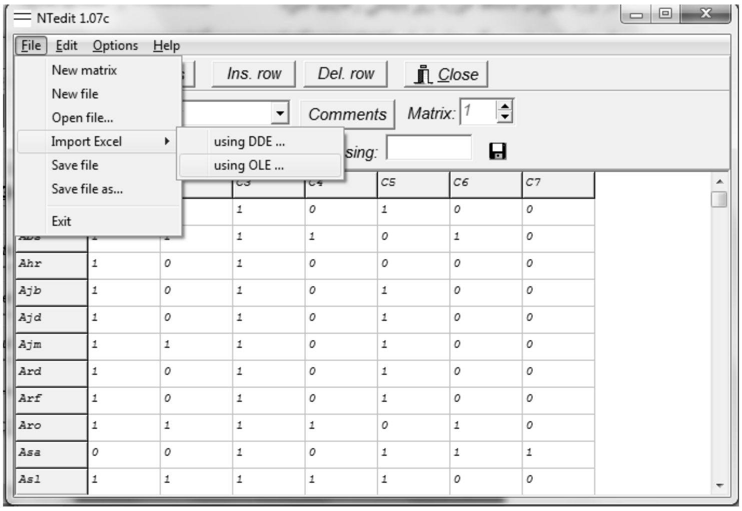


	A	B	C	D	E	F	G	H	I	J
1	1	250	7							
2										
3	Abk	0	0	1	0	1	0	0		
4	Abs	1	1	1	1	0	1	0		
5	Ahr	1	0	1	0	0	0	0		
6	Ajb	1	0	1	0	1	0	0		
7	Ajd	1	0	1	0	1	0	0		
8	Ajm	1	1	1	0	1	0	0		
9	Ard	1	0	1	0	1	0	0		
10	Arf	1	0	1	0	1	0	0		
11	Aro	1	1	1	1	0	1	0		
12	Asa	0	0	1	0	1	1	1		
13	Asl	1	1	1	1	1	0	0		
14	Azr	1	0	1	0	0	0	0		
15	Bnb	0	1	1	1	1	0	0		
16	Did	1	0	1	1	1	0	0		

Figure 1 – Preparation and data entry in Microsoft Excel 2010

Transfer saved data to NTedit 1.07c program in the following order:

- Open NTedit program.
- Click “File” menu then “Import Excel” sub-menu.
- Use “using DDE” or “using OLE” for opening .xls files in higher and lower versions of Microsoft Excel 2000, respectively.
- Recall the saved .xls file.
- Click “File” menu and select “Save file” sub-menu to save the file with arbitrary name with .NTS filename extension.
- Close NTedit program.



			C3	C4	C5	C6	C7
			1	0	1	0	0
			1	1	0	1	0
Ahr	1	0	1	0	0	0	0
Ajb	1	0	1	0	1	0	0
Ajd	1	0	1	0	1	0	0
Ajm	1	1	1	0	1	0	0
Ard	1	0	1	0	1	0	0
Arf	1	0	1	0	1	0	0
Aro	1	1	1	1	0	1	0
Asa	0	0	1	0	1	1	1
Asl	1	1	1	1	1	0	0

Figure 2 – Data transferring to NTedit 1.07c

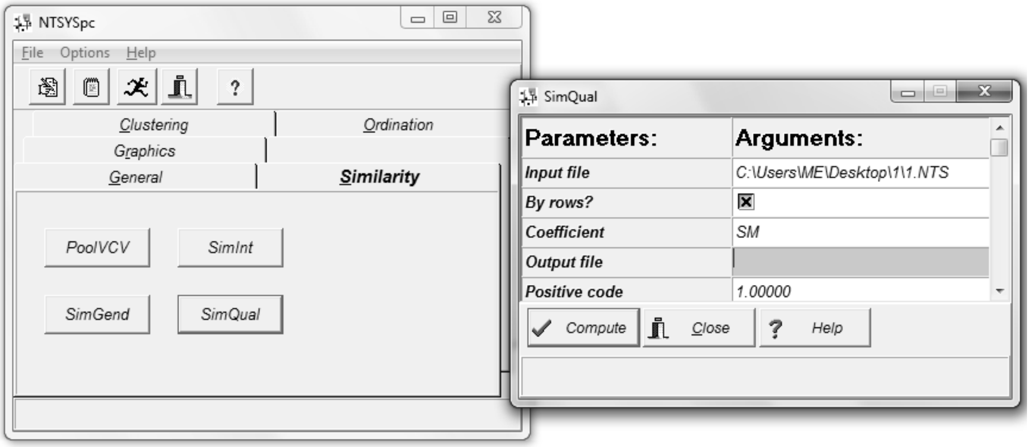
2.2. Clustering Analysis

This part is dendrograms' drawing procedure such as phylogenetic trees. First, you should open NTSYSpc ver. 2.02e program. This main part has the three conspicuous stages:

1. **Establishing similarity matrix:**
   - Go to "Similarity" tab.
   - Click on "SimQual" toolbar which is specific for making qualitative data like molecular's.
   - Recall the previous saved file with .NTS extension in "Input file" field.
   - Tick the box in front of "by rows" part.
   - Choose your desired method for making similarity matrix from the "Coefficient" part.

     ⚐ **Important**: "SM" method stands for "Simple Matching" is usually used for qualitative data including molecular's.

   - Specify the name and path of your output file in the filed "Output file".
   - Click on "Compute" toolbar.

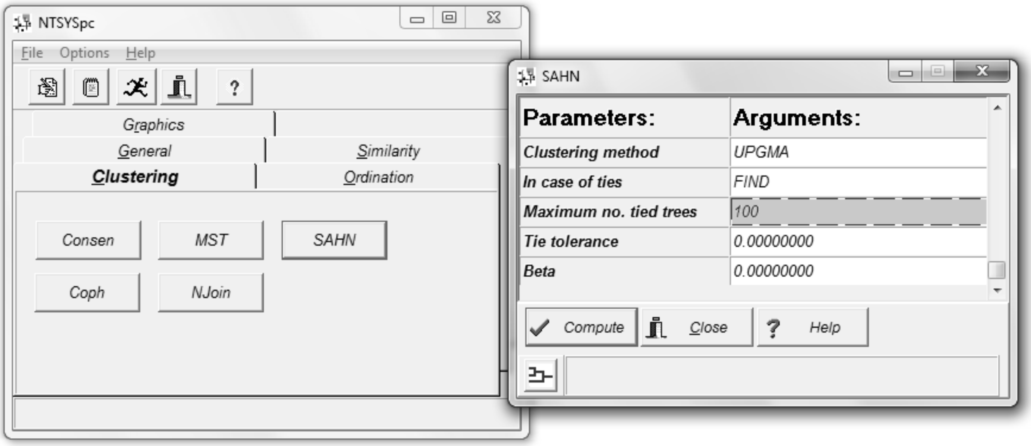


Parameters:	Arguments:
Input file	C:\Users\ME\Desktop\1\1.NTS
By rows?	☒
Coefficient	SM
Output file	
Positive code	1.00000

Figure 3 – Making similarity Matrix in NTSYSpc 2.02e

2. **Dendrogram drawing**
   - Go to "Clustering" tab.
   - Click on "SAHN" toolbar.
   - Recall the previous saved file (containing similarity matrix) in "Input file" field.
   - Specify the name and path of your output file in the filed "Output file".

- Choose desired method for clustering from the “Clustering method” part (such as UPGMA).
- Choose “FIND” option in the field of “In case of ties” part.
- If you have near 100 individuals, enter “100” in “Maximum no. ties tress” part.
- Do not change “Tie tolerance” and “Beta” fields.
- Click on “Compute” toolbar.
- Close the “Report listing” windows if the procedure was successful.
- Click on “pot tree” toolbar.
- A dendrogram will be drawn.



Parameters:	Arguments:
Clustering method	UPGMA
In case of ties	FIND
Maximum no. tied trees	100
Tie tolerance	0.00000000
Beta	0.00000000

Figure 4 – Dendrogram drawing by NTSYSpc 2.02e

⚐ **Important 1**: If you have more than 30 individuals, the software will show only up to 30 cases in dendrogram. For observing the other cases, double left click on dendrogram, the “Tree plot” dialog box will be opened. Go to “Page control” part and choose “OUTs/Page” and change it up to your case number.

⚐ **Important 2**: On “Tree plot” dialog box, in “Edit” menu there are two options. Choose “Copy metafile” for transferring diagram image with better resolution to clipboard and you can paste it on a Microsoft Office Word file.

3. Calculating and drawing of cut-off line

The mean of the similarity matrix data would be the cut off line position on dendrogram. Follow the following order to calculate it.

- Open similarity matrix data file that you have made in 2.1.1 part in Notepad program.
- Select all numbers (only numbers) in this file.
- Copy and paste the numbers in a Microsoft Office Word file.
- Open “Replace” dialog box using ‘Ctrl H” short-cut key and replace all spaces with tab, i.e. insert “space” in “Find what” and “^t” in “replace with” parts. Click “Replace all” to do all replacements
- Copy and paste these data to a blank Microsoft Excel file and calculate the average, using “average” in “Autosum” parts in “formulas” menu.
- The obtained number will be the position of cut-off line on dendrogram.
- Draw the cut-off line on dendrogram and specify how many main clades you have got.
- Count how many individuals you have in each clade.

2.3. Selection of individuals

For all primers you have got, you should do the above-mentioned procedure, separately. Now, try to insert the number of individuals that a given individual is common in the same clade in each primer’s dendrogram in the sample table below. Finally add up them in the last right column (Total column).

	Primer 1	Primer 2	…..	Primer n	Total

Individuals					
Individual 1					
Individual 2					
….					
Individual n					

Finally, sort the individuals in ascending order according to “Total” column in the Microsoft Office Word. The individuals will be arranged by the least unity with others based on all primers. Now, you can choose how many individuals you want for further analysis like DNA sequencing and so on.

2.4. Conclusions

Initial screening and individual selection is necessary in genetic diversity studies of huge populations. The screening can be done by some molecular techniques like RAPD and MSP-PCR. As DNA sequencing is an expensive method, we may screen and select some most vitiated cases for it. This will be resulted in saving time and cost in these sorts of investigations, in deed. Therefore, conscious selection of more unique individuals with the least common features, considering multiplicity of band patterns of molecular techniques, supposed to be the most important part of this process. Using this method, the individuals can be selected with high accuracy and based on the least relativeness with other individuals for further assessments. Additionally, the method can be generalized in other similar studies with slight modifications.

3. Acknowledgements

Special thanks to Dr. Davood Hasanpanah and Dr. Naser Mohebbalipour for their convenient aids.

4. References

[1] A.D. Baxevanis, B.F.F. Ouellette. Bioinformatics. A practical guide to the analysis of genes and proteins.. Wiley-Interscience. Third Edition. 2005, 560 pp.

[2] Microsoft Corporation. Microsoft Excel 2010 product guide. 2010, 76 pp.

[3] Microsoft Corporation. Microsoft Word 2010 product guide.2010, 66 pp.

[4] M. Parani, A. Anad, and A. Parida. Application of RAPD Fingerprinting in Selection of Micropropagated Plants of *Piper longum* for Conservation. *Current Science*. 1997, **73**(1): 81.

[5] F.J. Rohlf. NTSYSpc Numerical Taxonomy and Multivariate Analysis System Version 2.0 User Guide. Applied Biostatistics Inc., Setauket, New York. 1998, 37 pp.

[6] S. Spooner, van Treuren R, and M.C. de Vicente. Molecular markers for genebank management. CGN, IPGRI, USDA. 2005, 126 p.

[7] P. Sunnucks. Efficient Genetic Markers for Population Biology. Tree. 2005, **15**(5): 199-203.

[8] H. Tanaka, S. Sawairi, T. Okuda. Application of the random amplified polymorphic DNA using the polymerase chain reaction for efficient elimination of duplicate strains in microbial screening. III. Bacteria. J. Antibiot. 1994, **47**(2):194-200.

[9] A. TariNezhad, A. Sabouri, S.A., and Mohammadi. Statistical software NTSYS pc application in plant breeding. The 7th Conference of Iran Statistics. Allame Tabatabaei University, September 2005. Tehran, Iran. 10pp.

[10] Wang L, Dang Z, and Zhang J. Preliminary Screening of RAPD Markers for Fertility Gene of Thermo-sensitivity Male-sterile Line1S of Flax. Acta Agriculturae Boreali-Occidentalis Sinica. 2009, **1**(8): 123-127.

[11] Wikipedia. Screening. Retrieved from <http://en.wikipedia.org/wiki/Screening> on 19 July 2011.