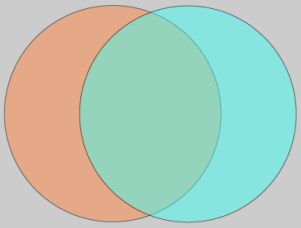
# SONS



## Introduction

SONS (**S**hared **O**TUs a**n**d **S**imilarity) is a computer program that uses non-parametric estimators to estimate similarity between communities based on membership and structure. Because SONS is directly compatible with output files from DOTUR, it is possible to quickly determine the fraction of OTUs shared by two communities for any desired distance level.

Briefly, SONS reads in a line from a DOTUR-formatted *.list file and the names and library designation for each sequence from a user-generated tab-delineated file (the same format as used in TreeClimber). SONS then determines the number individuals in each community that were sampled for each OTU. Next SONS calculates collector's curves for the fraction of shared OTUs between the two communities (with and without correcting for unsampled individuals), the Jaccard and Sorenson Indices, and the richness of OTUs shared between the two communities. Standard error values are calculated for entire sequence collection. SONS is freely available as C++ source code and as a Windows executable.

This manual is designed to achieve four goals:

1. Show how to use SONS
2. Describe output files and equations used to calculate each parameter
3. Validate output by making calculations by hand
4. Answer frequently asked questions

If you have any questions, complaints, or praise, please do not hesitate to contact Dr. Patrick D. Schloss at pschloss@microbio.umass.edu

## How to Run SONS

To compile SONS in LINUX type the following in the folder with the makefile:

>g++ sons.C –O4 –o sons

SONS is run from the command line prompt and requires two input files. You first need a DOTUR-formatted *.list file. The *.list file can be generated using DOTUR by generating a sequence alignment of all the sequences in the comparison, then using DNADIST from PHYLIP to generate a distance matrix, and then using DOTUR. Alternatively, you can make your own *.list file by following this format: each line begins with either a distance or a one-word name describing the comparison, whitespace, a number indicating the number of OTUs that are being considered, whitespace, and then the names of sequences in each OTU separated by commas. Each OTU is separated by a space. For example, the first line of 70.fn.list from the Eckburg study would look like this (NOTE: There are no returns after each line):

```
unique  2742
        B001.contigs,B034.contigs,B070.contigs,B071.contigs,B175.contigs,B191.contigs,B260.contig
s,B264.contigs,B334.contigs,B336.contigs,B350.contigs,B391.contigs,B510.contigs,B619.contigs,B645
.contigs,B647.contigs,B685.contigs,B783.contigs,B851.contigs,B883.contigs,B912.contigs,B919.conti
gs,B979.contigs,BA17.contigs,BA59.contigs,BB07.contigs      B003.contigs   B004.contigs
        B005.contigs   B006.contigs   B007.contigs   B008.contigs
        B009.contigs,B169.contigs,B277.contigs,B704.contigs,BA79.contigs,BB02.contigs
        B010.contigs   B011.contigs   B014.contigs
        B015.contigs,B876.contigs,BB59.contigs,G280.contigs (etc until there are descriptions of
all 2742 OTUs).
```

Next SONS requires a tab-delineated file containing the names of each sequence in the first column and the library designation in the second column. For example, the *.names file for a comparison of the stool and mucosal membership from patient 70 in the Eckburg study would look like this (see file: 70.stool_compare.names):

```
K003.contigs   tissue
K004.contigs   tissue
K005.contigs   tissue
K006.contigs   tissue
K008.contigs   tissue
K010.contigs   tissue
K011.contigs   tissue
K012.contigs   tissue
.
.
.
BB90.contigs   stool
BB91.contigs   stool
BB92.contigs   stool
BB93.contigs   stool
BB94.contigs   stool
BB95.contigs   stool
```

Since Eckburg, et al. sampled 4,392 sequences from patient 70, the names file would contain 4,392 rows and 2 columns. You can as many library names as you desire and SONS will calculate every possible pairwise comparison. A spreadsheet program such as Microsoft Excel or OpenOffice Calc is possibly the easiest way to generate the file. An important consideration is that he name of the sequences in the *.names file match, identically, the names used to generate the *.list file.

With the *.list file and *.names file in hand you are now ready to run SONS. In linux the following command will execute SONS using the default parameters:

```
>./sons -list 70.fn.list -names 70.stool_compare.names
```

Execution in Windows and Linux (and Mac OSX) is essentially the same. In Windows, you cannot merely double click on the icon to get the program to execute. You must use the "Command Prompt" program found by going Start -> Program Files -> Accessories -> Command Prompt. Then you must type in the path of SONS and your distance file to execute the program:

```
C:\> "Documents and Settings\pds\Desktop\sons.exe" -list "Documents and
Settings\pds\Desktop\70.fn.list" -names "Documents and
Settings\pds\Desktop\70.stool_compare.list"
```

Alternatively, you can change the root path to move to the desired directory and execute SONS from there:

```
C:\PATH\> sons.exe -list 70.fn.list -names 70.stool_compare.names
```

Be forewarned that SONS does not seem to run as quickly in Windows as it does in Linux and I would encourage everyone to align their sequences in ARB, which uses Linux or OSX, and to run DOTUR and SONS in the same operating system.

Once executed, the program will begin to churn and you will see the progress of the random iterations and some data for interpreting your results. Remember that because two pairwise comparisons are made for every comparison, it is essential that you correct for multiple comparisons.

Other options exist for running SONS. In default mode, SONS will use 1,000 iterations to calculate the standard error for each parameter. This can be changed by setting the –i flag as follows:

```
>./sons -list 70.fn.list -names 70.stool_compare.names -i 10000
```

Another option is to randomize the order of the sequences listed in the *.names file by setting the –jumble flag. The default is to construct collector's curves using the order of the sequences given in the *.names file.

```
>./sons -list 70.fn.list -names 70.stool_compare.names -jumble
```

These flags will work when running sons.exe in Windows the same as they are implemented in Linux.

## Output Files

SONS produces three output files: *.sons, *.sons.ltt, and *.sons.otu. I will explain what each file contains and then how the calculations were derived.

***.sons***

Data to construct collector's curves for each comparison and distance level are provided in the *.sons file. If you have a number of comparisons and sequences in your analysis, then this file can become quite unwieldy. I would suggest using either "grep" commands or perl scripts to parse the file into more manageable units. The columns are formatted as follows and the first row of the file contains the information listed in the second row of this table (the equations and sample calculations will be given later):

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Distance** | **A** | **B** | **Number of Seq. Sampled** | **$U_{hat}$** | **$V_{hat}$** | **$UV_{hat}$** | **Jabund** | **Labund** | **$V_{est}$** | **Shared Chao** | **$U_{obs}$** | **$V_{obs}$** | **AOTU Shared** | **BOTU Shared** | **$J_{class}$** | **$L_{class}$** | **thetaYC** | **thetaYC (se)** | **thetaN** |
| 0.03 | tissue | Stool | 4392 | 0.97 | 0.95 | 0.93 | 0.93 | 0.96 | 72.5 | 76.7 | 0.93 | 0.94 | 0.67 | 0.74 | 0.55 | 0.71 | 0.53 | 0.03 | 0.88 |

The data is sorted according to column 1, then 2, then 3, and then 4. This data was taken from the file "70.fn.sons."

***.sons.ltt***

The formatting of the *.sons.ltt file is essentially the same as the *.sons file except that column 4 is not provided and a standard error is calculated for the parameters in columns 5-10. The first line of the *.sons.ltt file has a description of each column's contents. Each row contains the estimates for the calculations performed using all of the appropriate data in the dataset. Again the data is sorted according to column 1, then 2, and then 3.

***.sons.otu***

This file contains the frequency of sequences from each library found in each OTU. The first row of each file contains the distance being considered so that columns with the same value in the first row go together. The second row tells which library the data represents and the third row indicates the number of sequences sampled from that library. Each subsequent row represents a different OTU so that the number indicates the number of sequences in that library that clustered within that OTU. Note that OTU frequencies can only be compared within a distance definition. Here is an example of data from the 70.fn.sons.otu file:

| **Distance Level** | 0.03 | 0.03 |
|---|---|---|
| **Library Designation** | tissue | stool |
| **Number of OTUs** | 110 | 110 |
| **Num seqs. in OTU 1** | 1135 | 174 |
| **Num seqs. in OTU 2** | 91 | 27 |
| **…** | | |
| **Num seqs. in OTU 110** | 1 | 0 |

## Example Calculations

Example calculations will be performed using data from the 70.sons.fn.otu file with an OTU definition of 0.03.

**Estimating the richness of shared OTUs between two communities.** Non-parametric richness estimators of the number of shared OTUs between two communities have been developed that are analogous to the Chao1 (1) and ACE (5) single community richness estimators. The $S_{A,B\ Chao}$ (6) and $S_{A,B\ ACE}$ (4) estimators are calculated as:

$$S_{A,B\ \boldsymbol{ACE}} = S_{\mathbf{12}(abund)} + \frac{S_{\mathbf{12}(rare)}}{C_{\mathbf{12}}} + \frac{\mathbf{1}}{C_{\mathbf{12}}}\left[f_{(rare)\mathbf{1}+}\Gamma_{\mathbf{1}} + f_{(rare)+\mathbf{1}}\Gamma_{\mathbf{2}} + f_{\mathbf{11}}\Gamma_{\mathbf{12}}\right] \qquad \text{(column 10)}$$

$$S_{A,B\ Chao} = S_{\mathbf{12}(Obs)} + f_{\mathbf{11}}\frac{f_{\mathbf{1}+}f_{+\mathbf{1}}}{\mathbf{4}f_{\mathbf{2}+}f_{+\mathbf{2}}} + \frac{f_{\mathbf{1}+}^{\mathbf{2}}}{\mathbf{2}f_{\mathbf{2}+}} + \frac{f_{+\mathbf{1}}^{\mathbf{2}}}{\mathbf{2}f_{+\mathbf{2}}} \qquad \text{(column 11)}$$

where,

$$C_{\mathbf{12}} = \mathbf{1} - \frac{\sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}}\{Y_i I(X_i = \mathbf{1}) + X_i I(Y_i = \mathbf{1}) - I(X_i = Y_i = \mathbf{1})\}}{\sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} X_i Y_i}$$

$$\Gamma_{\mathbf{1}} = \frac{S_{\mathbf{12}(rare)} n_{rare} T_{\mathbf{21}}}{C_{\mathbf{12}}(n_{rare} - \mathbf{1})T_{\mathbf{10}}T_{\mathbf{11}}} - \mathbf{1}, \quad \Gamma_{\mathbf{2}} = \frac{S_{\mathbf{12}(rare)} m_{rare} T_{\mathbf{12}}}{C_{\mathbf{12}}(m_{rare} - \mathbf{1})T_{\mathbf{01}}T_{\mathbf{11}}} - \mathbf{1}$$

$$\Gamma_{\mathbf{3}} = \left[\frac{S_{\mathbf{12}(rare)}}{C_{\mathbf{12}}}\right]^{\mathbf{2}} \frac{n_{rare} m_{rare} T_{\mathbf{22}}}{(n_{rare} - \mathbf{1})(m_{rare} - \mathbf{1})T_{\mathbf{10}}T_{\mathbf{01}}T_{\mathbf{11}}} - \frac{S_{\mathbf{12}(rare)} T_{\mathbf{11}}}{C_{\mathbf{12}}T_{\mathbf{01}}T_{\mathbf{10}}} - \Gamma_{\mathbf{1}} - \Gamma_{\mathbf{2}}$$

$$T_{\mathbf{10}} = \sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} X_i\ ,\ T_{\mathbf{01}} = \sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} Y_i\ ,\ T_{\mathbf{11}} = \sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} X_i Y_i\ ,\ T_{\mathbf{21}} = \sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} X_i(X_i - \mathbf{1})Y_i\ ,$$

$$T_{\mathbf{12}} = \sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} X_i(Y_i - \mathbf{1})Y_i\ ,\ T_{\mathbf{22}} = \sum_{i=\mathbf{1}}^{S_{\mathbf{12}(rare)}} X_i(X_i - \mathbf{1})Y_i(Y_i - \mathbf{1})$$

$f_{11}$ = number of shared OTUs with one observed individual in A and B

$f_{1+}$, $f_{2+}$ = number of shared OTUs with one or two individuals observed in A

$f_{+1}$, $f_{+2}$ = number of shared OTUs with one or two individuals observed in B

$f_{(rare)1+}$ = number of OTUs with one individual found in A and less than or equal to 10 in B.

$f_{(rare)+1}$ = number of OTUs with one individual found in B and less than or equal to 10 in A.

$n_{rare}$ = number of sequences from A that contain less than 10 sequences.

$m_{rare}$ = number of sequences from B that contain less than 10 sequences.

$S_{12(rare)}$ = number of shared OTUs where both of the communities are represented by less than or equal to 10 sequences

$S_{12(abund)}$ = number of shared OTUs where at least one of the communities is represented by more than 10 sequences

$S_{12(Obs)}$ = number of shared OTUs in A and B

Calculation of column 11 requires the number of OTUs where only one sequence was observed from each library, $f_{11}$. For our example case, $f_{11}$ is 2. Plugging the f-values and $D_{12}$ into equation 9 yields a value of 76.7, which matches the value in column 11 of the table above.

Calculation of column 10 is considerably more complicated to evaluate. First, we determine that there are 23 rare shared OTUs and 37 abundant shared OTUs. Next, considering only the rare OTUs, we calculate $C_{12}$ as 0.845878. We obtained the following T-values:

| | |
|---|---|
| $\mathbf{T_{10}}$ | 93 |
| $\mathbf{T_{01}}$ | 64 |
| $\mathbf{T_{11}}$ | 279 |
| $\mathbf{T_{21}}$ | 1444 |
| $\mathbf{T_{12}}$ | 988 |
| $\mathbf{T_{22}}$ | 5440 |

Next, calculation of the Γ-values requires knowing $f_{(rare)1+}$, $f_{(rare)+1}$, and $f_{(rare)11}$, which were 5, 8, and 2. Also, $n_{rare}$ and $m_{rare}$ were 185 and 167, respectively. Finally, calculation of the Γ-values gives $\Gamma_1$=0.530409, $\Gamma_2$=0.523308, and $\Gamma_3$=0.151840. This gives a $S_{A,B\ ACE}$ (equation 10) of 72.5 as reported in column 10 above.

**Estimating the fraction of shared OTUs between two communities.** Incidence-based measures of community similarity such as the classic Jaccard ($J_{clas}$) and Sørenson ($S_{clas}$) similarity indices calculate the ratio of shared OTUs to the total number of OTUs in individual communities:

$$J_{clas} = \frac{S_{12}}{S_1 + S_2 - S_{12}} \qquad \text{(column 16)}$$

$$S_{clas} = \frac{2S_{12}}{S_1 + S_2} \qquad \text{(column 17)}$$

where,

$S_1$, $S_2$ = number of OTUs observed or estimated in A and B
$S_{12}$=number of OTUs shared between A and B

The observed number of OTUs in A and B was 89 and 81, respectively. Shared between the two libraries were 60 OTUs. Therefore the value of the equations for columns 16 and 17 were 0.55 and 0.71, respectively. An alternative expression is the fraction of observed OTUs that were shared (columns 14 and 15):

$$A_{shared} = \frac{S_{12}}{S_1} \qquad \text{(column 14)}$$

$$B_{shared} = \frac{S_{12}}{S_2} \qquad \text{(column 15)}$$

The values for these were 0.67 and 0.74, respectively.

**Estimating the fraction of sequences that belong to shared OTUs.** Just as the Chao1 richness estimator is a function of the number of OTUs observed once or twice in a sample (1), the

estimators of the fraction of sequences in shared OTUs is a function of the number of shared OTUs that are observed at least once or twice in the community being analyzed (3, 2):

$$U_{est} = \sum_{i=1}^{D_{12}} \frac{X_i}{n_{total}} + \frac{m_{total} - 1}{m_{total}} \frac{f_{+1}}{2f_{+2}} \sum_{i=1}^{D_{12}} \frac{X_i}{n_{total}} I(Y_i = 1) \quad \text{(column 5)}$$

$$V_{est} = \sum_{i=1}^{D_{12}} \frac{Y_i}{m_{total}} + \frac{n_{total} - 1}{n_{total}} \frac{f_{1+}}{2f_{2+}} \sum_{i=1}^{D_{12}} \frac{Y_i}{m_{total}} I(X_i = 1) \quad \text{(column 6)}$$

where,

$U_{est}$, $V_{est}$ = fraction of sequences from A and B that belong to a shared OTU
$X_i$, $Y_i$ = abundance of the $i^{th}$ shared OTU in A and B
$n_{total}$, $m_{total}$ = total number of sequences sampled in A and B
I(•) = if the argument, •, is true then I(•) is 1; otherwise it is 0.

For this example, $n_{total}$ and $m_{total}$ were 3,332 and 1,060, respectively and the first summation in equations 4 and 5 equal 0.927971 and 0.94434, respectively. $f_{1+}$, $f_{2+}$ equal 5 and 3, respectively and $f_{+1}$, $f_{+2}$ equal 15 and 10, respectively. The second summations in equations 4 and 5 equal 0.057323 and 0.010377, respectively. Evaluating equations 4 and 5 then gives 0.97 and 0.95 as shown in columns 5 and 6 of the table above. Note that these values are considerably greater than those derived from equations 1 and 2. The 95% confidence intervals for equations 5-6 can be determined by a bootstrapping method described elsewhere (2).

$U_{obs}$ and $V_{obs}$ (columns 12 and 13) represent the number of fraction of sequences in A and B respectively that are found in shared OTUs:

$$U_{obs} = \frac{m_{shared}}{m_{total}} \quad \text{(column 12)}$$

$$V_{obs} = \frac{n_{shared}}{n_{total}} \quad \text{(column 13)}$$

In our case the number of shared sequences in A and B ($m_{shared}$ and $n_{shared}$) was 3,092 and 1,001. Therefore, the resulting values in column 12 and 13 were 0.93 and 0.94, respectively.

Using these estimators, the abundance-based Jaccard ($J_{abund}$) and Sørenson ($S_{abund}$) similarity indices may be calculated (3, 2):

$$J_{abund} = \frac{U_{est} V_{est}}{U_{est} + V_{est} - U_{est} V_{est}} \quad \text{(column 8)}$$

$$S_{abund} = \frac{2U_{est} V_{est}}{U_{est} + V_{est}} \quad \text{(column 9)}$$

Evaluation of the equations for columns 8 and 9 yield 0.93 and 0.96, respectively. These were the values reported in column 8 and 9 in the table above. The 95% confidence intervals for these equations can be determined by a bootstrapping method described elsewhere (2).

**Estimating community structure similarity.** The overlap measures described by the equations for columns 8 and 9 do not account for the similarity of the relative abundances among the OTUs

shared between two communities. Therefore, although they measure community overlap, they do not measure the similarity of two community structures. Yue and Clayton (7) proposed a non-parametric maximum likelihood estimator of similarity, θ, to compare community structures:

$$\theta_{YC} = \frac{\sum_{i=1}^{S_{12}} \frac{X_i}{n_{total}} \frac{Y_i}{m_{total}}}{\sum_{i=1}^{S_1} \left( \frac{X_i}{n_{total}} \right)^2 + \sum_{i=1}^{S_2} \left( \frac{Y_i}{m_{total}} \right)^2 - \sum_{i=1}^{S_{12}} \frac{X_i}{n_{total}} \frac{Y_i}{m_{total}}} \qquad \text{(column 18)}$$

The 95% confidence intervals be determined using the explicit variance formula for θ that was derived by Yue and Clayton (7) and is presented in column 19. By plugging in the values for each parameter here, the value of θ for column 18 is 0.53 and the standard error in column 19 is 0.03.

Another similarity measure of community structure is (7):

$$\theta_N = \frac{\sum_{i=1}^{S_{12}} \frac{X_i}{n_{total}} \sum_{i=1}^{S_{12}} \frac{Y_i}{m_{total}}}{\sum_{i=1}^{S_{12}} \frac{X_i}{n_{total}} + \sum_{i=1}^{S_{22}} \frac{Y_i}{m_{total}} - \sum_{i=1}^{S_{12}} \frac{X_i}{n_{total}} \sum_{i=1}^{S_{12}} \frac{Y_i}{m_{total}}} \qquad \text{(column 20)}$$

This is essentially a form of the equation for column 8 that is not corrected for the presence of unsampled OTUs. For this example the value was 0.88.

## Frequently Asked Questions

***How do I cite SONS?***
Schloss, P.D. & Handelsman, J. 2006. Introducing SONS, a tool for OTU-based comparisons of membership and structure between microbial communities. *Applied and Environmental Microbiology*. In review.

***In windows, why does the command window open and close quickly when I double click on the SONS icon in windows?***
This is because you have not given SONS the input files and you will get an error message quickly followed by the screen closing. Please see the above section on how to run SONS and remember that it must be run from a command prompt in windows.

***Why doesn't SONS do…?***
If you would like to see something added to SONS, please let me know (pschloss@microbio.umass.edu). It may take me a while to get around to implementing the feature, but I am generally reasonable.

## References


1. **Chao, A.** 1984. Non-parametric estimation of the number of classes in a population. Scand. J. Stat. **11:**265-270.
2. **Chao, A., R. L. Chazdon, R. K. Colwell, and T. J. Shen.** 2006. Abundance-based similarity indices when there estimation when there are unseen species in samples. Biometrics **62:**361-371.
3. **Chao, A., R. L. Chazdon, R. K. Colwell, and T. J. Shen.** 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol. Lett. **8:**148-159.
4. **Chao, A., W. H. Hwang, Y. C. Chen, and C. Y. Kuo.** 2000. Estimating the number of shared species in two communities. Stat. Sinica. **10:**227-246.
5. **Chao, A., and S. M. Lee.** 1992. Estimating the number of classes via sample coverage. J. Am. Stat. Assoc. **87:**210-217.
6. **Chao, A., T. J. Shen, and W. H. Hwang.** 2006. The applications of Laplace's boundary-mode approximations to estimate species richness and shared species richness. Aust. N. Z. J. Stat. **48:**117-128.
7. **Yue, J. C., and M. K. Clayton.** 2005. A similarity measure based on species proportions. Commun. Stat. Theor. M. **34:**2123-2131.