

# BIONUMERICS Tutorial: Creating a custom mappings similarity matrix

# 1 Aim

In BIONUMERICS, character values can be mapped to categorical names according to predefined criteria (see tutorial "Importing non-numerical character data" for more information about the use of mappings in BIONUMERICS). When character mappings are present, it becomes possible to define a custom mappings similarity matrix, which determines how similarities are calculated among the mappings. This can be useful when analyzing data sets like SNPs, VNTRs, SSRs, etc. In this tutorial the use of a custom mappings matrix is illustrated.

# 2 Preparing the database

#### 2.1 Introduction to the SNP demonstration database

The **SNP demonstration database** contains SNP data for 37 entries (see Figure 1) and can be downloaded directly from the *BIONUMERICS Startup* window (see 2.2), or restored from the back-up file available on our website (see 2.3).

10 SNPs were screened, and since the screening was performed on diploid organisms, 10 states (4 homozygous and 6 heterozygous states) are possible: A:A, C:C, G:G, T:T, C:A, G:A, T:A, G:C, T:C, and T:G.

### 2.2 Option 1: Download demo database from the Startup Screen

1. Click the **button**, located in the toolbar in the *BIONUMERICS Startup* window.

This calls the *Tutorial databases* window (see Figure 2).

- 2. Select the SNP demonstration database from the list and select *Database* > *Download* (2).
- 3. Confirm the installation of the database and press < OK > after successful installation of the database.
- 4. Close the *Tutorial databases* window with *File* > *Exit*.

SNP demonstration database - BioNumerics		– 🗆 X
File Edit Database Analysis Scripts Window Help		
☞ 글   ▦   ☞   ♡   ኴ,		
Experiment types	Database entries	Comparisons
୶ + r≯ ⊗ B <sub>4</sub>   B₁	# + P ⊗ €   E マ. <all entries=""></all>	+ 🖻 🛞 🗟   🛱 🤜 «All Comparisons»
# Name Type 🖵	Key Comment 🖵 1	Name Modified date Level 🖵
1 SNP calls Character types	Strain01	
	Strain02	
	Strain03	
v	Strain04	< >>
	Strain05	
Entry fields Database design	Strain06	Identification projects Decision networks
	Strain07 •	29 + PA ⊗ B   Ph マ. <all identification<="" td=""></all>
	Strain08 •	
Name Field type	Strain09 •	Name Modified date
Comment Fixed	Strain10 •	
	Strain11 •	
v	Strain 12 •	
	Strain13 •	
Fingerprint files Power assemblies Annotations	Strain14 •	Alignments BLAST projects Chromosome comparisons
	Strain15 •	
	Strain 16 •	
File name Experiment type Link 🔻	Strain17 •	Name Modified date 🗸
	Strain18 •	
	Strain 19 •	
	Strain20 •	
< >>	□ Strain21 • V	
Database: SNP demonstration database (_DefaultUser_) Entries: Loaded=37	7, View=37, Selected=0 1 experiments C:\Users\Public\Documents\BioNumerics\Data BN8	SNP demonstration database This is a time limited package valid until 2020-

Figure 1: The Main window of the SNP demo database.

🖆 Tutorial databases					– 🗆 X
File Database Tutorial Window					
r Tutorial databases					
1 T					
Name	Size (MB)	Min	Max	•	
BandScoring demo database	15	7	8	~	The DemoBase Connected demonstration
DemoBase Connected	4	7.5	8		database in BioNumerics version 8.0 format.
Neisseria MLST demo database	1	7.5	8		Inis database is used in several tutorials and
SNP demonstration database	0	7.5	8		comes with the Diolyumencs installation.
WGS demo database for Listeria monocytogenes	311	7.5	8		
WGS_demo_database_for_Brucella_spp	214	7.6	8		
WGS_demo_database_for_Escherichia_coli	601	7.5	8		
WGS_demo_database_for_Staphylococcus_aureus	624	7.5	8		
	_			~	×
	<			>	
Tutorials					
Ē					
Name				-	
Band matching and polymorphism analysis				~	This tutorial illustrates how to calculate a
Building automated decision and action workflows					Principal Components Analysis (PCA) and a
Calculating a PCA and an MDS on a fingerprint data set					(sometimes also called Principal
Clustering a phenotypic test assay					Coordinates Analysis (PCoA)) on a
Combined analysis of character data					fingerprint data set and how to change the
Combined analysis of fingerprint data					layout of the obtained plots.
Configuring the database layout					
Dendrogram layout options					V
Entry information fields and their properties	_			~	



#### The SNP demonstration database appears in the BIONUMERICS Startup window.

5. Double-click the **SNP demonstration database** in the *BIONUMERICS Startup* window to open the database.

The Main window should look like Figure 1.

#### 2.3 Option 2: Restore demo database from back-up file

A BIONUMERICS back-up file of the SNP demo database is also available on our website. This backup can be restored to a functional database in BIONUMERICS.

6. Download the file SNPs.bnbk from https://www.applied-maths.com/download/sample-data, under "SNP demonstration database".



In contrast to other browsers, some versions of Internet Explorer rename the SNPs.bnbk database backup file into SNPs.zip. If this happens, you should manually remove the .zip file extension and replace with .bnbk. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the .zip file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

- 7. In the *BIONUMERICS Startup* window, press the button. From the menu that appears, select **Restore database...**.
- 8. Browse for the downloaded file and select *Create copy*. Note that, if *Overwrite* is selected, an existing database will be overwritten.
- 9. Specify a new name for this demonstration database, e.g. "SNP demonstration database".
- 10. Click < OK > to start restoring the database from the backup file.
- 11. Once the process is complete, click < Yes > to open the database.

The *Main* window should look like Figure 1.

# 3 Character mappings

1. Double-click on the **SNP calls** character type experiment in the *Experiment types* panel of the *Main* window to open the *Character type* window.

The SNP set is displayed in the *Characters* panel (see Figure 3). For each SNP a character range is specified from 0 to 10, corresponding to the 10 different states (4 homozygous states, 6 heterozygous states).

2. Click on the *Mapping* panel to display the character mappings defined (see Figure 4).

The 10 possible states are mapped: A:A, C:C, G:G, T:T, C:A, G:A, T:A, G:C, T:C, and T:G.

- 3. Close the *Character type* window.
- 4. Click on a green colored dot in the *Experiment presence* panel to open the **SNP calls** experiment card for an entry (see Figure 5 for an example).
- 5. Close the experiment card by clicking in the left upper corner of the card.

۵	haracter type 'SNP calls	5'				-		×
File	Settings Characters	Fields Mapping	Window Help					
E	 							
E	_ + ↑ + <u>└</u> ⊆ + 1010							
Cha	iracters							
				15			_	
		MX   1 +		0				
	Character	Enabled	Min. Max.	Color scale				
	SNP01	×	0	10				^
	SNP02	×	0	10				
	SNP03	× .	0	10				
	SNP04	× .	0	10				
	SNP05	× .	0	10				
	SNP06	× .	0	10				
	SNP07		0	10				
	SNP08		0	10				
	SNP09		0	10				
	SNP10	· · · · ·	0	10				
								~
Cha	racters Mapping							
Co	nparison settings							
c	D calls softings							^
		upp placed data ant	(10 shorestore)					
5	ve calls, numerical val	ues, closed data set	(TO CHARACTERS)					
- Comparison								
	Similarity coefficient							$\sim$
Con	Comparison settings Crosslinks Attachments							
Chara	atas tura SND aalla Ch	erectore in view-10						



🖆 Cha	racter ty	/pe 'SNP calls'					_	×
File Se	ttings	Characters I	Fields	Mapping	Window	Help		
ß	⊥†-	ite i™						
	1.11							
Маррі	ng							
Name		Range	e start		Range end			
No Call		0			0			
A:A		1			1			
C:A		2			2			
G:A		3			3			
C:C		5			4 5			
G:C		6			6			
T:C		7			7			
G:G		8			8			
T:G		9			9			
T:T		10			10			
Channe								
Charac	ters M	apping						
Comp	arison s	settings						
SNP	calls s	settings						^
SNP	calls: n	umerical valu	es, clos	sed data s	et (10 char	acters)		
			.,			,		
	- C	omparison						
		– Similar	ity co	efficient				~
Compa	rison seti	tings Crosslin	ks At	tachments				
01 1					_	_		



Strain26		
Character	Value	Mapping 🗨
SNP01	1	A:A
SNP02	1	A:A
SNP03	1	A:A
SNP04	3	G:A
SNP05	3	G:A
SNP06	3	G:A
SNP07	8	G:G
SNP08	1	A:A
SNP09	1	A:A
SNP10	3	G:A

Figure 5: The character experiment card.

### 4 Custom mappings similarity matrix

When character mappings are present, it becomes possible to define a custom mappings similarity matrix, which determines how similarities are calculated among the mappings.

- 1. Double-click on the **SNP calls** character type experiment in the *Experiment types* panel of the *Main* window to open the *Character type* window.
- Select Mapping > Edit mapping similarity matrix... to display the Mappings similarity matrix dialog box (see Figure 6).



Figure 6: Custom similarity matrix.

The *Mappings similarity matrix* dialog box shows all defined character mappings in a matrix format. The default similarity value is "1" or 100% match for self-matches (on the diagonal) and "0" for matches between different mappings. With other words, the default corresponds to a normal categorical matching. All values in the matrix, except for the ones on the diagonal, can be edited by clicking on the cell. Any value that deviates from the default "0" will be highlighted in green. In the custom similarity matrix of this database, the states with a difference in only one allele, have a 50% match ("0.5") (see Figure 6). The mappings similarity matrix will be used when similarities are calculated with the *Categorical (mappings)* coefficient in the *Comparison* window (see 6).

3. Close the *Mappings similarity matrix* dialog box and the *Character type* window.

## 5 Comparison window

- 1. Make sure the *Database entries* panel is the active panel, and select all entries with the keyboard shortcut **Ctrl+A**.
- Highlight the *Comparisons* panel in the *Main* window and select *Edit* > *Create new object...* (+) to create a new comparison for the selected entries.

All 37 entries are loaded in the Comparison window.

- 3. Click on the onext to the experiment name **SNP calls** in the *Experiments* panel to display the data in the *Experiment data* panel.
- 4. To show the character values as colors according to the color scale defined for each character, select *Characters* > *Show values* (III).
- 5. To display the mapped values, select *Characters* > *Show mappings* (III).
- 6. Select *Characters* > *Show mappings+colors* (I to show the mappings in overlay with the colors (see Figure 7).



Figure 7: The Comparison window.

### 6 Cluster analysis

Cluster analysis is a two-step process. First, all pairwise similarity values are calculated with a **similarity coefficient**. Then, the resulting similarity matrix is converted into a dendrogram with a **clustering algorithm**. Although in practice these steps are performed together, they each require their own comparison settings.

1. Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)...* in the *Comparison* window.

The first step deals with the similarity coefficient for the calculation of the similarity matrix.

2. Select *Categorical (mappings)* from the list and press <*Next*> (see Figure 8).

Comparison settings						
Page 1 Similarity coefficient						
Keep existing similarity matrix   Image: Pearson correlation   <	] Calculate as distance ] Ignore zero values					
	< Back	Next >	Canc	el		

Figure 8: Cluster analysis: step 1.

The mappings similarity matrix will be used when similarities are calculated with the *Categorical* (*mappings*) coefficient.

In step two the options related to the clustering algorithms are grouped. Under *Method*, the clustering algorithm to be applied on the similarity matrix can be selected. A *Dendrogram name* can be entered in the corresponding text box. By default, the name of the experiment type appended with the aspect (here: "SNP calls (<All characters>)") will be used.

3. Select *UPGMA* and *<Finish>* to start the cluster analysis.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels. The cluster analysis is listed in the *Analyses* panel of the *Comparison* window (see Figure 9).

4. Save the comparison with the dendrogram by selecting *File* > *Save* (ℍ, Ctrl+S). Specify a name (e.g. All) and press < *OK* >.



Figure 9: The Comparison window.