# Introduction to Sequence Analysis in R to create typologies of longitudinal trajectories

## Transcript video 2

Full resource (and video), see: <https://www.ncrm.ac.uk/resources/online/all/?id=20853>

Ian Thomas: In the previous video we created the sequences and in this one we’re going to compare them, and the reason we’re doing this is that the aim of typology creation is to actually look for patterns of similarities between the sequences, so in order to achieve that though we need to create a numerical representation of these similarities/the differences between sequences, and the way we achieve this is essentially by comparing each sequence to all of the other sequences within our sequence object.

 And the way this works is that we essentially take a sequence and try and change it into another sequence, and the number of times we have to change things within a sequence to make it look like another one essentially gives us an indication of how different they are, so if you’re having to change a lot of states within a sequence then the two things are obviously quite different.

 And because we’re comparing each sequence to all other sequences then this essentially generates what’s known as a distance matrix because we’re making pairwise comparisons for every sequence within our sequence object.

 So as with the creation of the sequences themselves it’s actually just a single line of code, the function that we’re calling in is called seqdist, so it’s the sequence distance, and essentially what this function is saying is draw on the sequence object, called in this case sequence object, and we’re going to use the method called HAM to look up the hamming distance between sequences.

 Hamming is just one example of how you can calculate the difference between a pair of sequences and it essentially operates by saying how many states within one sequence do I have to substitute with a different state in order to make two sequences match up?

 And as with anything in TraMineR, there are options about how you can calculate the cost attributed to making one of those substitutions, so you can choose to have a fixed cost so every time you change one element within a sequence that results in a cost of one, or you can base the costs on what’s known as the transition rates, so essentially this is saying if you’re trying to transform something into something that is less common then you’re obviously having to change a sequence in ways that are highly improbably and so that should deserve a higher cost.

 And we do this or we say we’re telling this function to use the transition rates to assign costs by using this SM equals trait option, and if I run that. So as before, a new object has appeared called distance matrix and the console gives you a summary of what happened when we told it to run that function. And again, it’s useful to look at this to make sure things have gone correctly, so we had 2,000 people in our biofam dataset and it’s still saying there’s 2,000 sequences and from our previous analysis there were eight distinct states and there are still eight distinct states.

 We told it to calculate the substitutions costs using the transition rates and it looks like it has done and there are still 16 unique elements within our states, which is again what we expect.

 And as before, I’m just going to open up the distance matrix so you can see what this function actually does. So like I said, it’s a pairwise matrix, so each person’s sequence is compared to all others, so person 1167 is being compared to the sequence for person 514 and then this is the overall cost or the overall value that has been attributed to all the different changes that have been made between those sequences in order to make them align.

 And because it’s a pairwise comparison then the diagonal will always have zeros because this is comparing person 116 to person 116, and quite obviously they are the same person so there’s no difference between their sequences.

 Some things to bear in mind, obviously this is a matrix that’s 2,000 columns by 2,000 rows because it’s a pairwise comparison, if your dataset is incredibly large then it will take obviously longer than 0.55 seconds. Equally, the length of your sequences can increase the length of time that it takes to construct this matrix.

 So that’s creating… well, that’s how we create this numerical representation of how different each sequence is to all others in the dataset and in the next video we’re going to actually conduct some cluster analysis of this distance matrix.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK