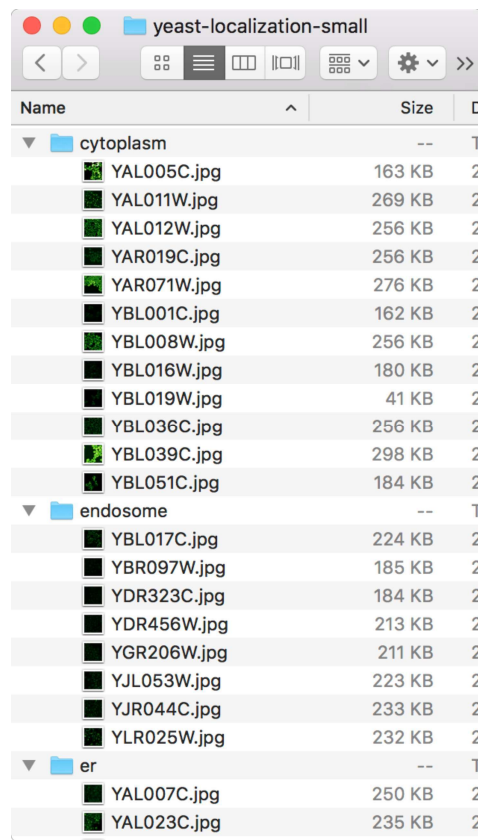


Lesson 36: Images and Classification

In this lesson, we are using images of yeast protein localization (<http://file.biolab.si/files/yeast-localization-small.zip>) in the classification setup. But this same data set could be explored in clustering as well. The workflow would be the same as the one from previous lesson. Try it out! Do Italian cities cluster next to American or are

We can use image data for classification. For that, we need to associate every image with the class label. The easiest way to do this is by storing images of different classes in different folders. Take, for instance, images of yeast protein localization. Screenshot of the file names shows we have stored them on the disk.



Name	Size	D
cytoplasm	--	Ti
YAL005C.jpg	163 KB	2
YAL011W.jpg	269 KB	2
YAL012W.jpg	256 KB	2
YAR019C.jpg	256 KB	2
YAR071W.jpg	276 KB	2
YBL001C.jpg	162 KB	2
YBL008W.jpg	256 KB	2
YBL016W.jpg	180 KB	2
YBL019W.jpg	41 KB	2
YBL036C.jpg	256 KB	2
YBL039C.jpg	298 KB	2
YBL051C.jpg	184 KB	2
endosome	--	Ti
YBL017C.jpg	224 KB	2
YBR097W.jpg	185 KB	2
YDR323C.jpg	184 KB	2
YDR456W.jpg	213 KB	2
YGR206W.jpg	211 KB	2
YJL053W.jpg	223 KB	2
YJR044C.jpg	233 KB	2
YLR025W.jpg	232 KB	2
er	--	Ti
YAL007C.jpg	250 KB	2
YAL023C.jpg	235 KB	2

Localization sites (cytoplasm, endosome, endoplasmic reticulum) will now become class labels for the images. We are just a step away from testing if logistic regression can classify images to their corresponding protein localization sites. The data set is small: you may use leave-one-out for evaluation in Test & Score widget instead of cross validation.

At about 0.9 the AUC score is quite high, and we can check where the mistakes are made and visualize these in an Image Viewer.

