Building scales for a Protein dataset

Albu Alexandra

ICA, group 246

Overview

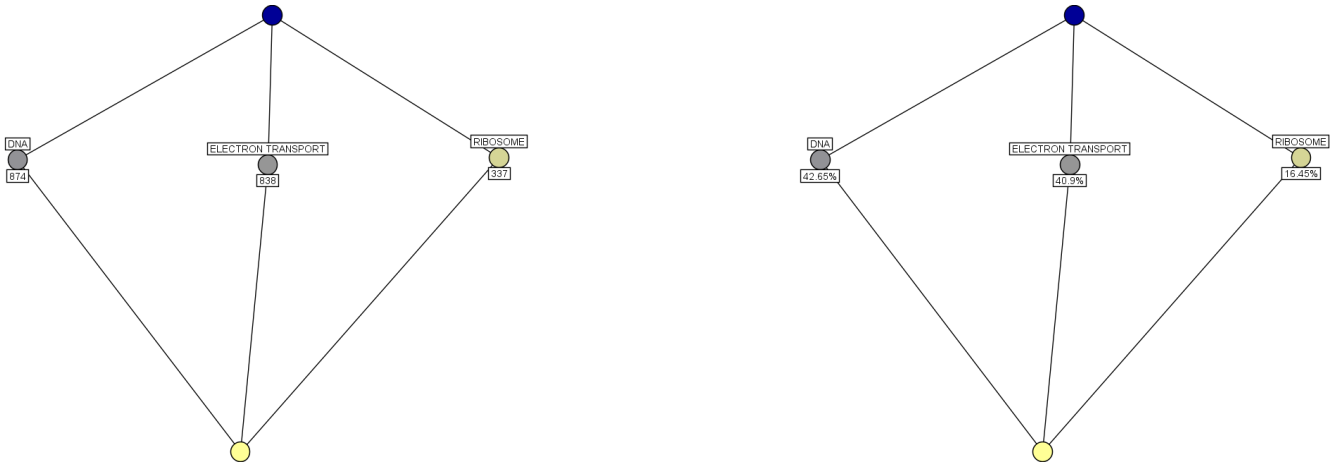
Attributes

Attribute	Type of values	Meaning
ID	real	protein ID
resolution	real	?
density (Matthews)	real	density of the crystal
Density (% solvent)	real	density computed from crystal cell dimensions and content
ph	real	pH value at which the crystal was grown
molecular weight	real	weight of a molecule
residue count	int	number of residues in the chain
macromolecule type	string	type of the molecule
experimental technique	string	experimental technique used to determine class
classification	string	the protein class

Table

Classification

- classification - categorical value: use a nominal scale



(a) Number of proteins per class (b) Distribution of proteins classes

Figure: Nominal scale for proteins classes

Residue Count

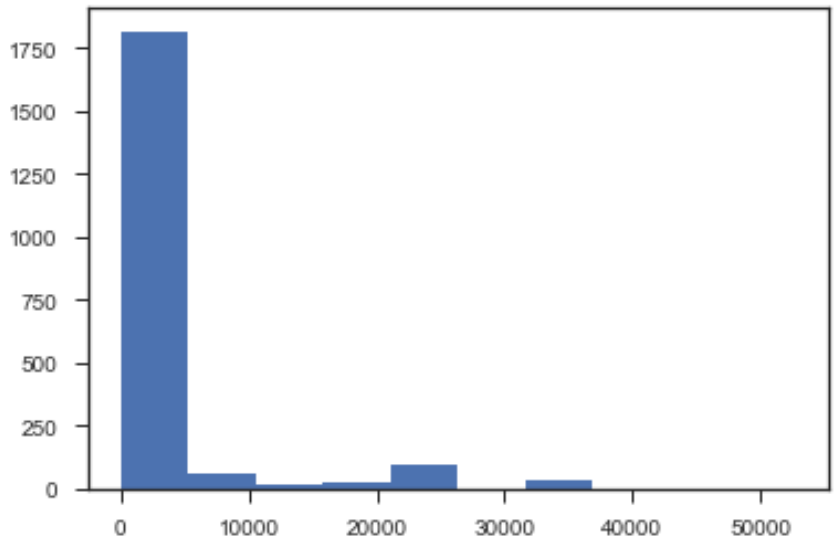


Figure: Histogram for Residue Count

- Observe that there are few proteins with more than 10.000 residues
- Question: is this property specific related to the class of the protein?

Residue Count for classes

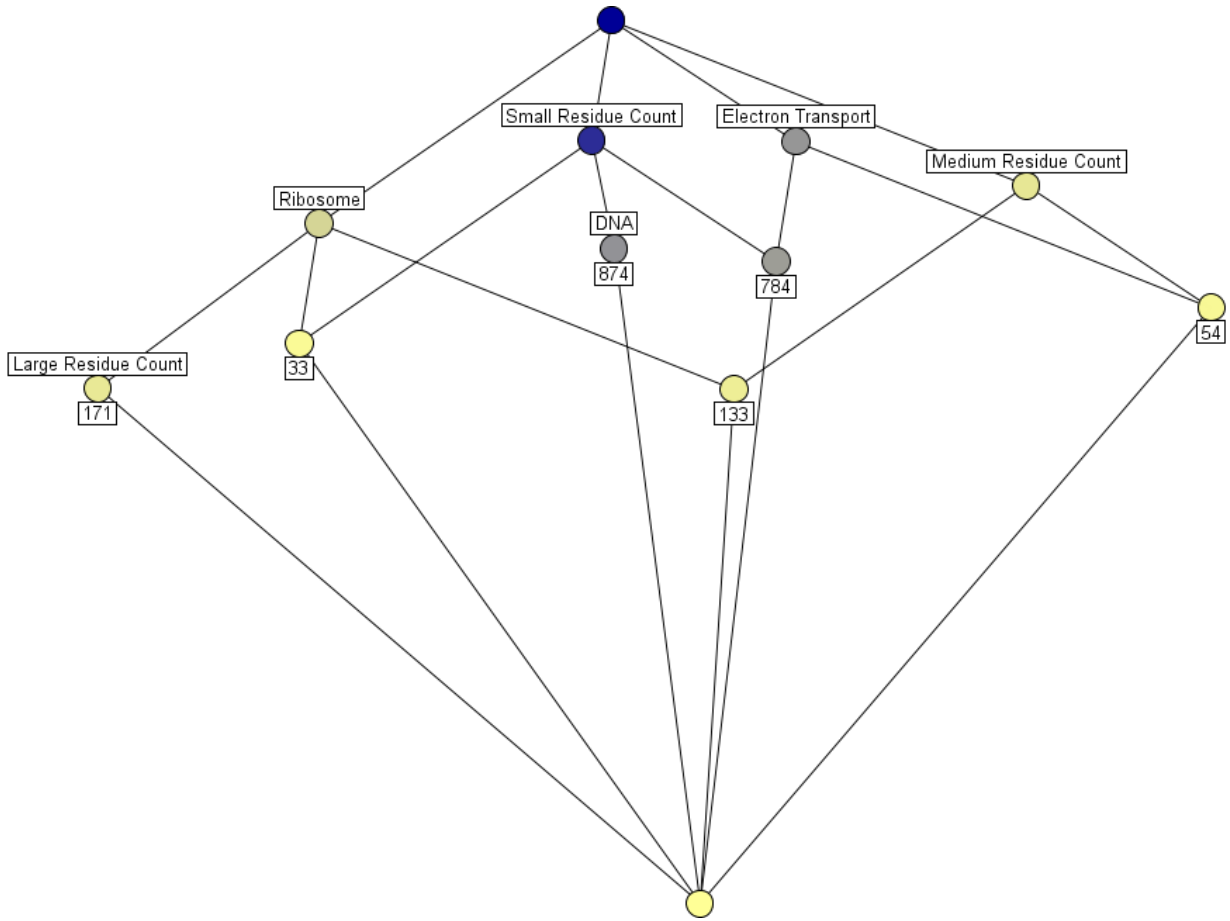
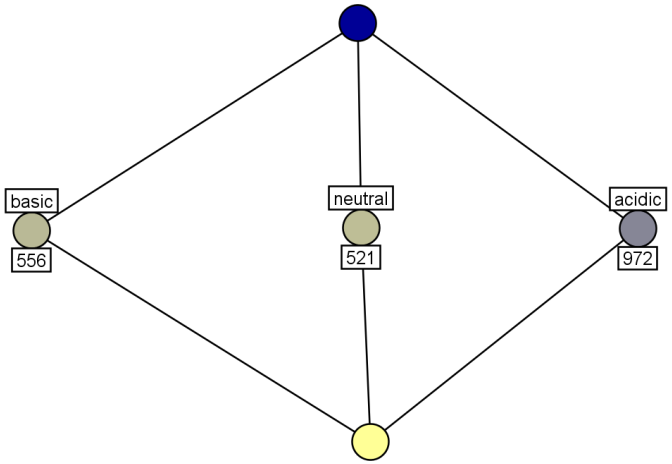
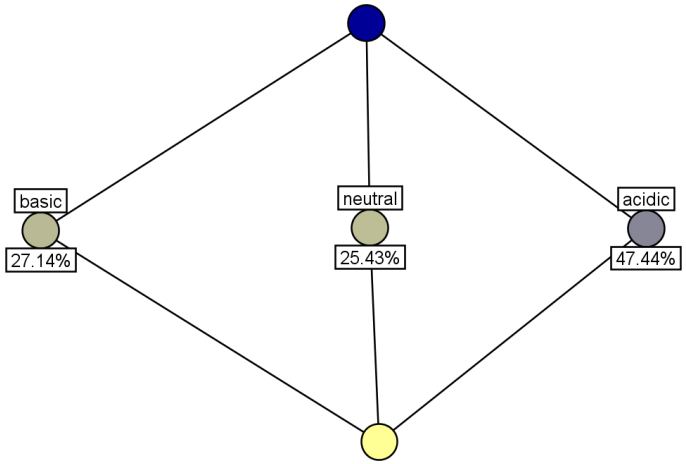


Figure: We create the following classes: Large Residue Count if the residueCount $>$ 10.000, Medium if in [1000, 10.000], small if $<$ 1000

pH



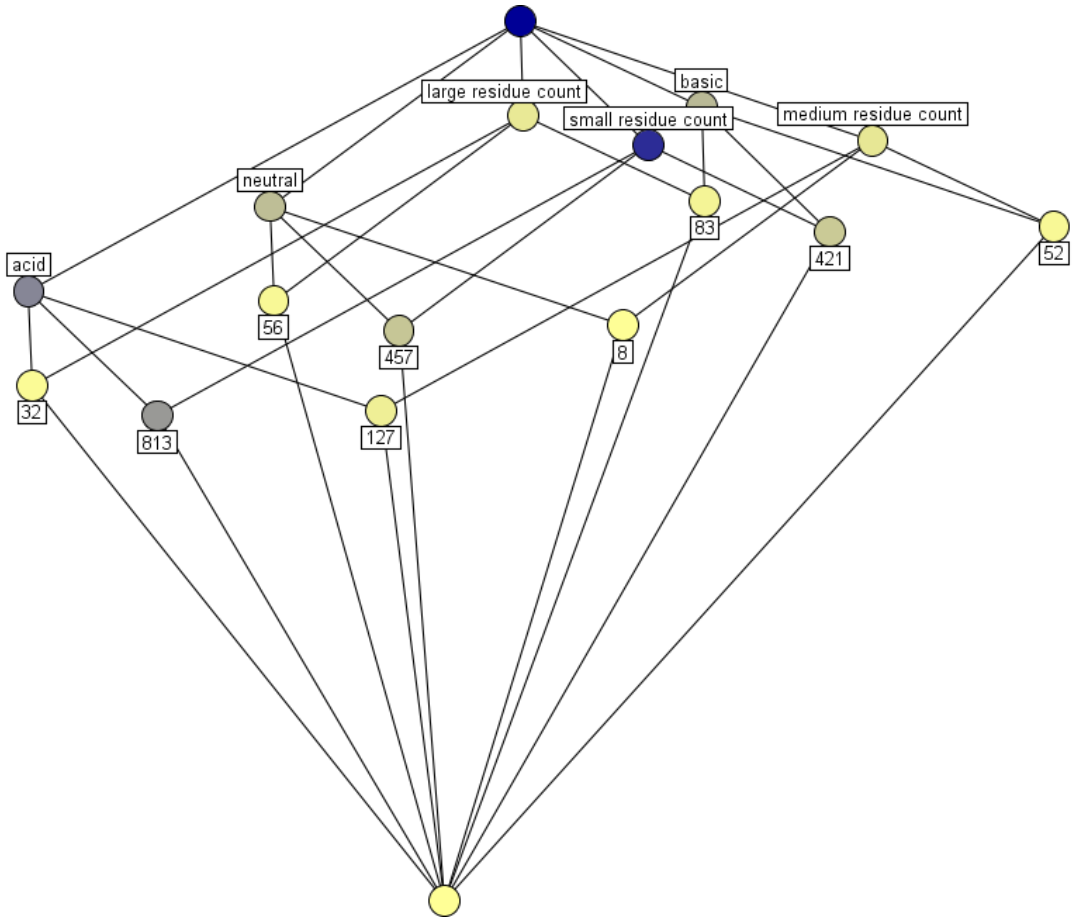
(a) Number of proteins per acidity class



(b) Distribution of acidity classes

Figure: Nominal scale for proteins classes

Residue Count for Acidity Classes



Figure

Resolution

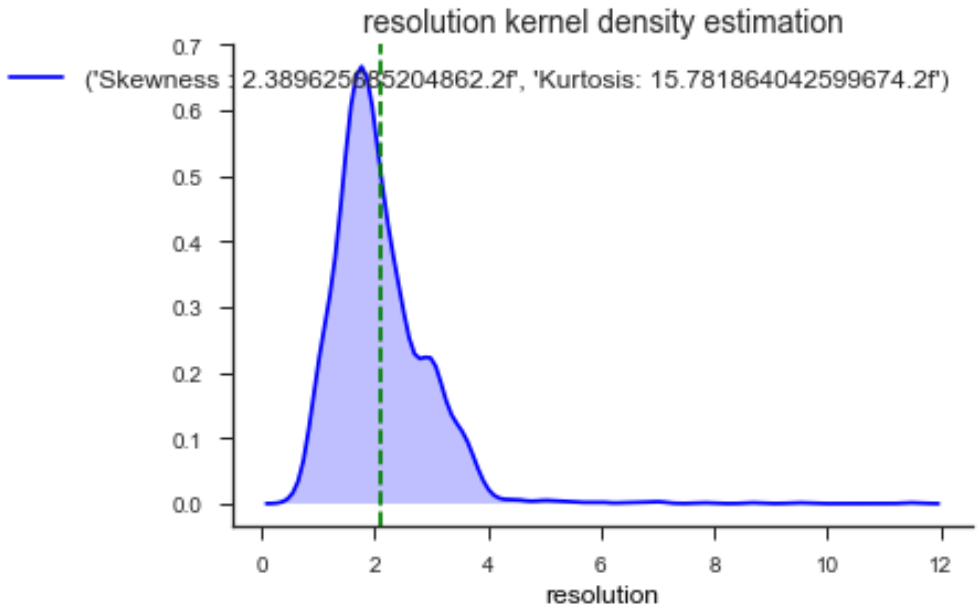


Figure: Caption

Conclusions

- data processing needed in order to discover knowledge
- visualization of attribute values (histograms, density estimations) - helpful in building meaningful scales