Problem 1:

1. The Bayes assumption probably does not hold, because GC content and complexity might have some dependence. This is because if GC content is 0 or 1, then the complexity will be low, but if GC content is around 0.5, the complexity is likely higher. Another example is that given that you have a coding region in a gene, you know that length is directly correlated with high GC content.
2. P(Y): P(Gene)=3/10, P(Repeat)=3/10, P(Motif)=4/10

P(GC=low|Gene) = 2/3, P(GC=medium|Gene)=1/3, P(GC=high|Gene)=0

P(GC=low|Repeat) = 0, P(GC=medium| Repeat)=0, P(GC=high| Repeat)=1

P(GC=low|Motif) = 0, P(GC=medium| Motif)=1/2, P(GC=high| Motif)=1/2

P(Length=long|Gene)=1, P(Length=short|Gene)=0

P(Length=long| Repeat)=1, P(Length=short| Repeat)=0

P(Length=long| Motif)=0, P(Length=short| Motif)=1

P(Complexity=high|Gene)=2/3, P(Complexity|low |Gene)=1/3

P(Complexity=high| Repeat)=1/3, P(Complexity|low | Repeat)=2/3

P(Complexity=high| Motif)=3/4, P(Complexity|low | Motif)=1/4

Problem 2:

1. See notebook
2. Coding to use the probability distribution of N and C to create sequences of scores

Problem 3:

1. Implemented
2. Plots submitted