**Course: Advanced Bioinformatics**

**Module title: Alphabets, Symbols and Sequences**

**Module no. : 193**

Symbols & Alphabets

* In BioJava the DNA residue “A” is an object.
* In Bioperl “A” would be a String.
* The “A” object is part of the sequence not the sequence.
* “A” from DNA is not equal to “A” from RNA or “A” from Protein.

Why not strings?

* DNA A != RNA A != Protein A
* For Strings “A”.equals(“A”);
* DNA Alphabet also contains K,Y,W,S,R,M,B,D,G,V,N
* Object Y contains C and T, The String “Y” doesn’t contain anything
* Translation HashMaps with Strings are flawed.
  + Biojava GGN translates to GLY
  + String GGN maps to null
* A fully redundant String to String HashMap translation table requires 4096 keys!

Symbols are Canonical

* DNATools.a() == DNATools.a();
  + There is only one instance of ‘a’
* DNATools.a().equals(DNATools.a());
* ProteinTools.a() != DNATools.a();
* Even on Remote JVM’s!
  + During serialization Alphabet indexing is transient and ‘reconnected’ via readResolve() methods.

Alphabets

* A set of Symbols
* Alphabets can be infinite
  + DoubleAlphabet, IntegerAlphabet
* Some Alphabets have a Finite number of Symbols
  + DNA, RNA etc
* Alphabet and FiniteAlphabet interfaces

org.biojava.bio.Alphabet

boolean [**contains**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)([Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) s)   
Returns whether or not this Alphabet contains the symbol.

List [**getAlphabets**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)()   
Return an ordered List of the alphabets which make up a compound alphabet.

[Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) [**getAmbiguity**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)(java.util.Set syms)   
Get a symbol that represents the set of symbols in syms.

[Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) [**getGapSymbol**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)()   
Get the 'gap' ambiguity symbol that is most appropriate for this alphabet

String [**getName**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)()   
Get the name of the alphabet.

[Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) [**getSymbol**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)(java.util.List rl)   
Get a symbol from the Alphabet which corresponds to the specified ordered list of symbols.

[SymbolTokenization](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\seq\io\SymbolTokenization.html) [**getTokenization**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)(java.lang.String name)   
Get a SymbolTokenization by name.

void [**validate**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Alphabet.html)([Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) s)   
Throws a precanned IllegalSymbolException if the symbol is not contained within this Alphabet.

void [**addSymbol**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\FiniteAlphabet.html)([Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) s)   
Adds a symbol to this Alphabet

Iterator [**iterator**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\FiniteAlphabet.html)()   
Retrieve an Iterator over the Symbols in this Alphabet.

void [**removeSymbol**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\FiniteAlphabet.html)([Symbol](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\Symbol.html) s)   
Remove a symbol from this alphabet.

int [**size**](file:///\\PHSGRI-S3002\schrema6$\src\cvs\biojava-live\doc\org\biojava\bio\symbol\FiniteAlphabet.html)()   
The number of symbols in the alphabet.

Default Alphabets

* DNA (a,c,g,t)
* RNA (a,c,g,u)
* PROTEIN (all amino acids including ‘Sel’)
* PROTEIN-TERM (all PROTEIN plus “\*”)
* STRUCTURE (PDB structure symbols)
* Alphabet of all integers (Infinite Alphabet)
  + Can generate SubIntegerAlphabets
* Alphabet of all doubles (Infinite Alphabet)

Getting the common Alphabets

**Import** org.biojava.bio.symbol.\*;  
 **import** java.util.\*;  
 **import** org.biojava.bio.seq.\*;  
   
 **public** **class** AlphabetExample {  
   **public** **static** **void** main(String[] args) {  
     Alphabet dna, rna, prot;  
   
     //get the DNA alphabet by name  
     dna = AlphabetManager.alphabetForName("DNA");  
   
     //get the RNA alphabet by name  
     rna = AlphabetManager.alphabetForName("RNA");  
   
     //get the Protein alphabet by name  
     prot = AlphabetManager.alphabetForName("PROTEIN");  
     //get the protein alphabet that includes the \* termination Symbol  
     prot = AlphabetManager.alphabetForName("PROTEIN-TERM");  
   
     //get those same Alphabets from the Tools classes  
     dna = DNATools.getDNA();  
     rna = RNATools.getRNA();  
     prot = ProteinTools.getAlphabet();  
     //or the one with the \* symbol  
     prot = ProteinTools.getTAlphabet();  
   
   }  
 }