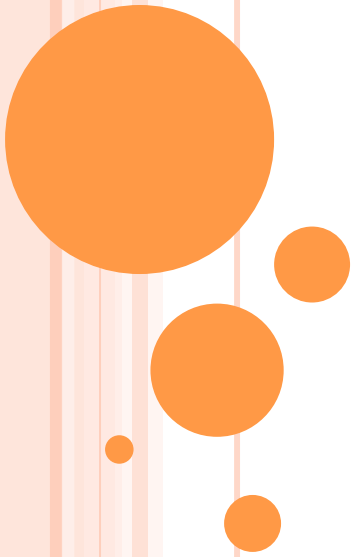


# ISF Overview

Vishal Nayak

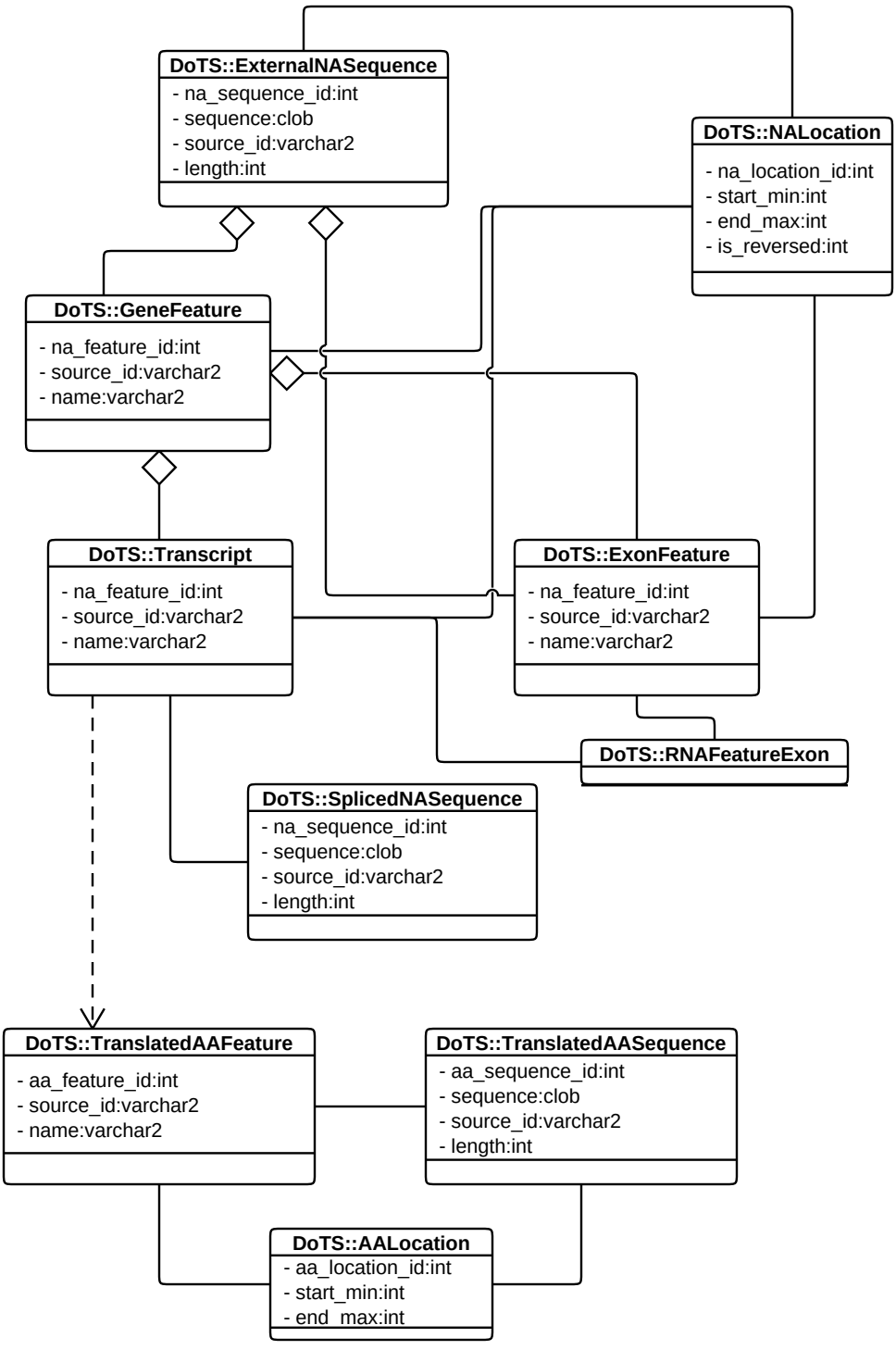
# WHAT IS ISF?

- ISF = InsertSequenceFeatures
- GUS Plugin to load features located on sequences.
- Designed to deal with
  - Different types of sequence features and feature trees
  - multiple file formats
  - Inconsistency across a single file format:
    - feature names
    - attributes
    - naming conventions
  - many-to-many mappings between features/ attributes and GUS tables/columns

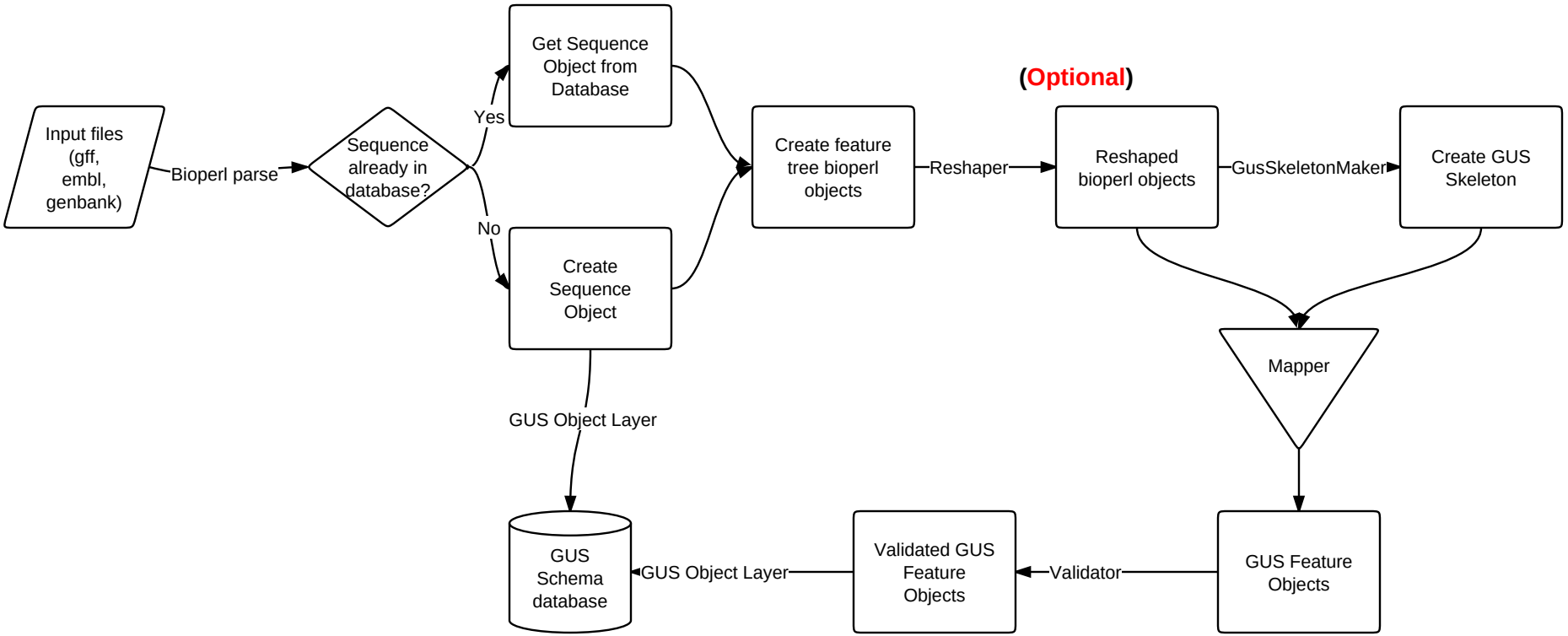
# What is ISF (continued)

- Supported file formats:
  - genbank, embl, gff2, gff3
- Loads:
  - Features on pre-existing sequences.
  - Or sequences and features simultaneously,
- Examples:
  - gene models, ORFs, isolates, binding site features.

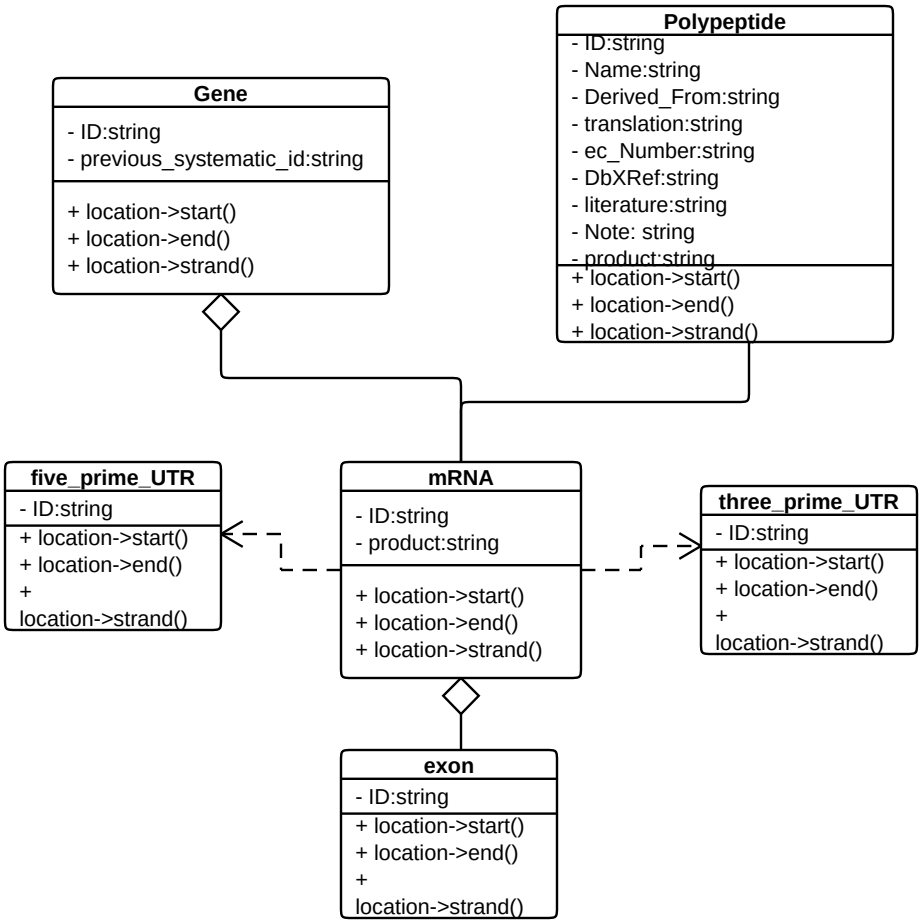
# Gene Models In GUS Schema



# ISF Conceptual Flowchart

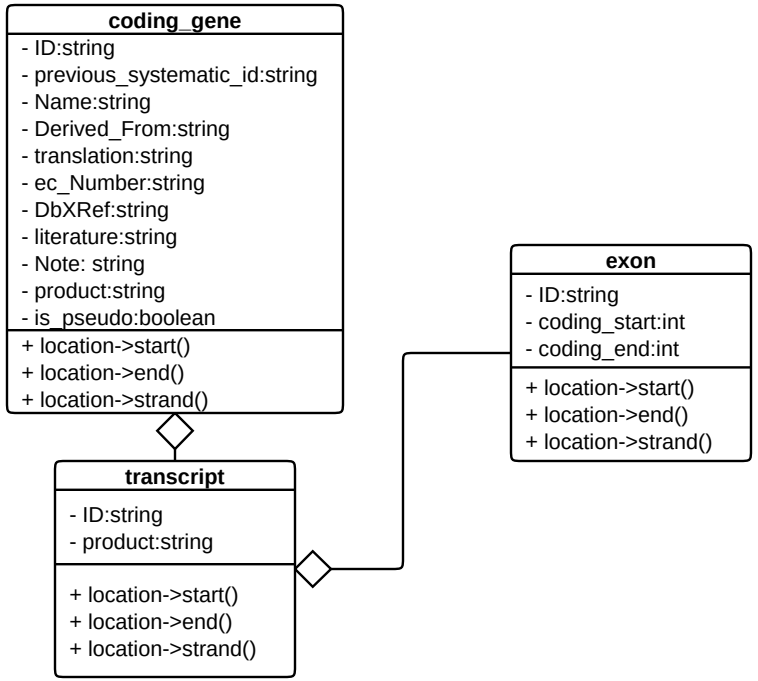


## Sanger GFF Gene Models (BioPerl)

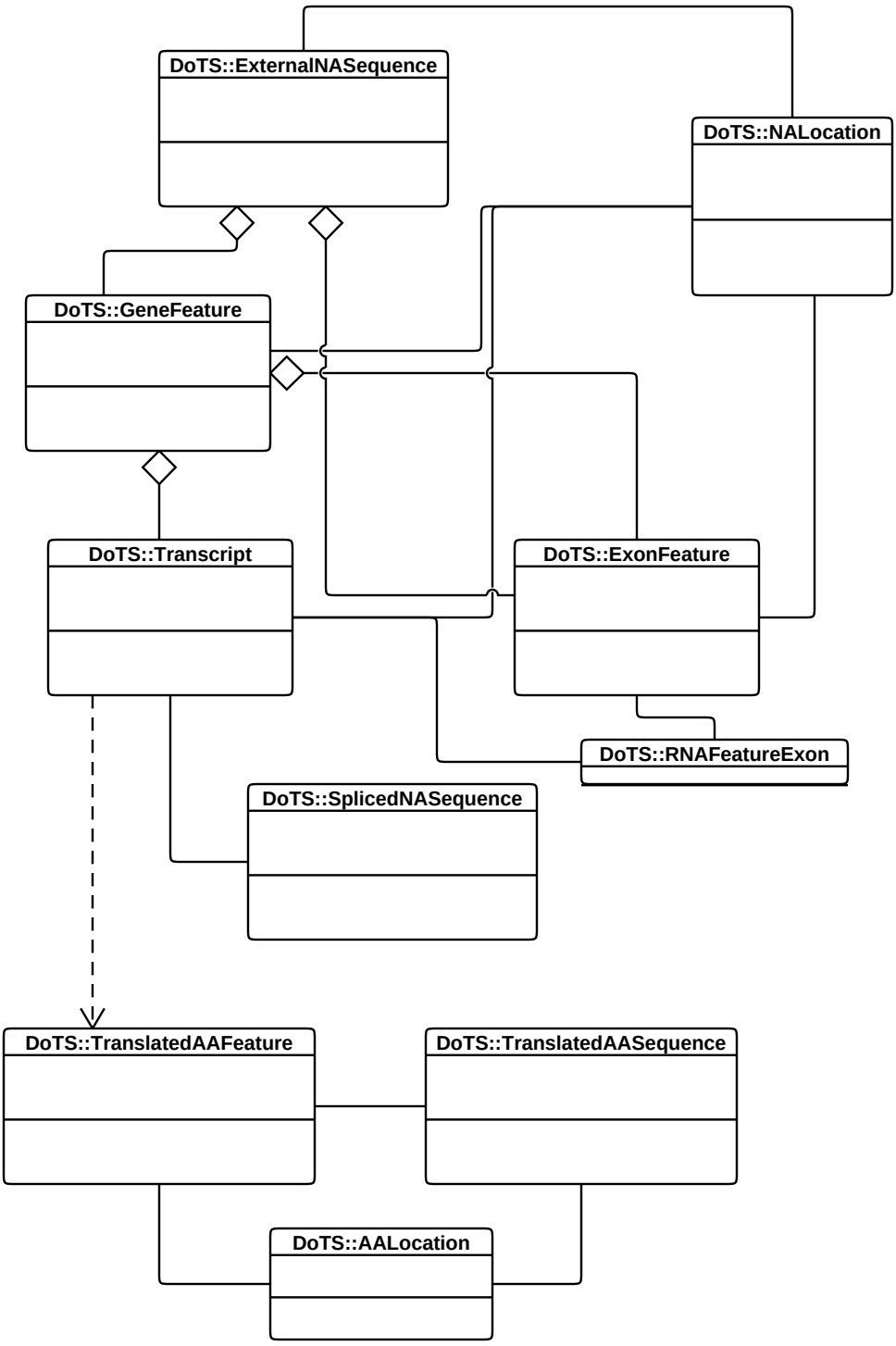



Reshaper

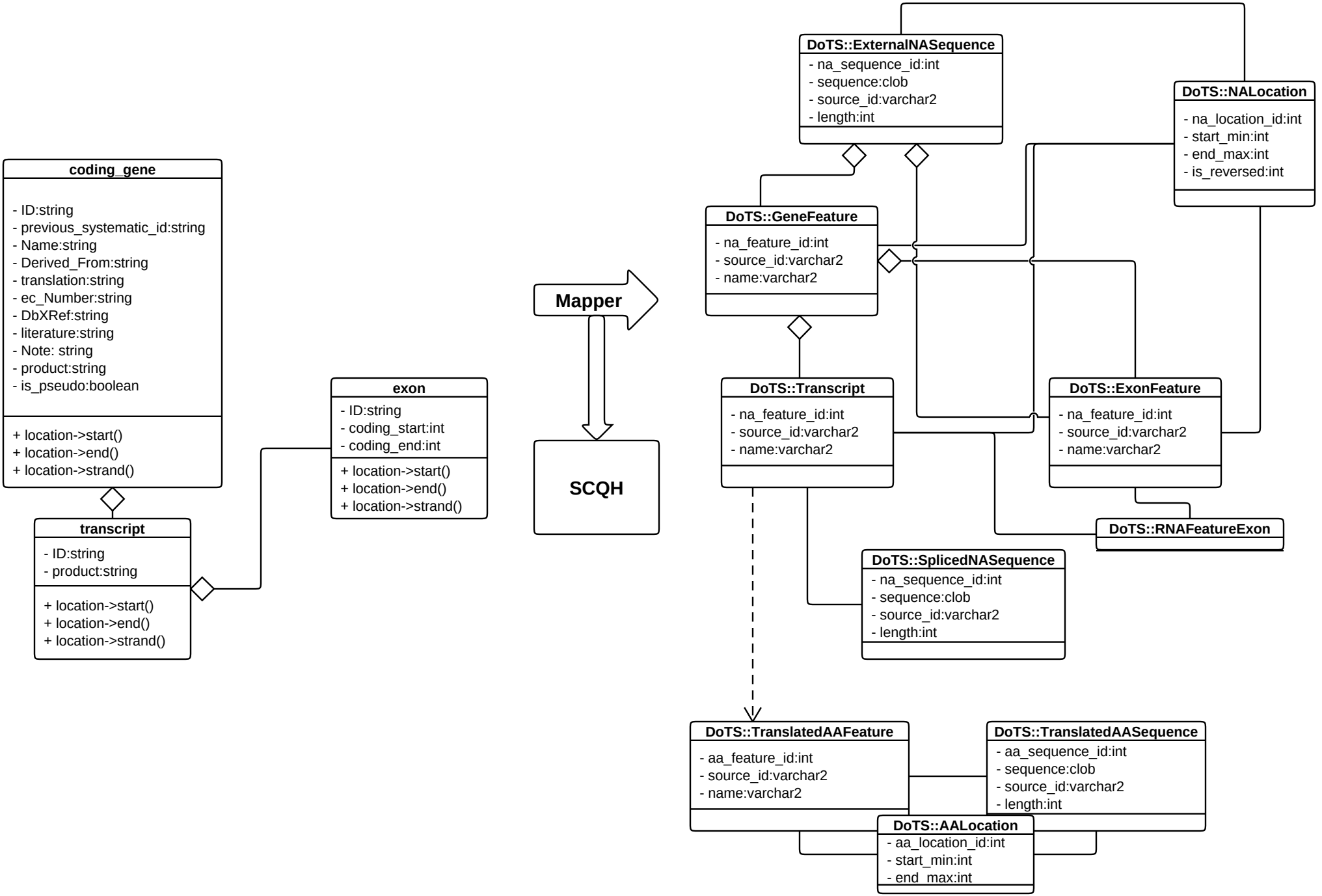
## Reshaped Gene Model (BioPerl)



# GUSSkeletonMaker - Gene Model Skeleton



# ISF Mapper


coding_gene
- ID:string
- previous_systematic_id:string
- Name:string
- Derived_From:string
- translation:string
- ec_Number:string
- DbXRef:string
- literature:string
- Note: string
- product:string
- is_pseudo:boolean
+ location->start()
+ location->end()
+ location->strand()
transcript
- ID:string
- product:string
+ location->start()
+ location->end()
+ location->strand()
exon
- ID:string
- coding_start:int
- coding_end:int
+ location->start()
+ location->end()
+ location->strand()
Mapper
SCQH
DoTS::ExternalNASequence
- na_sequence_id:int
- sequence:clob
- source_id:varchar2
- length:int
DoTS::NALocation
- na_location_id:int
- start_min:int
- end_max:int
- is_reversed:int
DoTS::GeneFeature
- na_feature_id:int
- source_id:varchar2
- name:varchar2
DoTS::Transcript
- na_feature_id:int
- source_id:varchar2
- name:varchar2
DoTS::ExonFeature
- na_feature_id:int
- source_id:varchar2
- name:varchar2
DoTS::RNAFeatureExon
DoTS::SplicedNASequence
- na_sequence_id:int
- sequence:clob
- source_id:varchar2
- length:int
DoTS::TranslatedAAFeature
- aa_feature_id:int
- source_id:varchar2
- name:varchar2
DoTS::TranslatedAASequence
- aa_sequence_id:int
- sequence:clob
- source_id:varchar2
- length:int
DoTS::AALocation
- aa_location_id:int
- start_min:int
- end_max:int

# Mapper

- An xml file that is required by ISF
- Maps:
  - Reshaped bioperl objects to GUS objects created by GUSSkeletonMaker.
  - Qualifiers/attributes to corresponding GUS objects/ attributes.
- Mapping between reshaped features/qualifiers and GUS objects/attributes might be many-to-many.
- If mapping between attributes is not one-to-one, we need **Special Case Qualifier Handlers**.

# ISF VALIDATION

- Validation is done on feature level (currently, only for coding genes).
- For coding genes:
  - Calculates protein translation and matches it against the provided translation. If translations are not provided, uses the calculated translations.
  - Checks if said gene has a source id.
  - Issues a warning if the sequence contains stop codons in the middle.
- For features requiring validation, a <validator> tag has to be specified in the mapper.xml file.

# ISF LIMITATIONS

- Currently unable to deal with gff3 format files directly.
  - Sequences need to be moved to a fasta file and loaded separately.
  - Separate gff3 files need to be merged and preprocessed into a “pseudo gff3” format.
- We are not yet dealing with alternative splicing.
- The validator uses an utility for calculating protein translations that requires all three bases in a codon to identify an amino acid.