From Ontology for Genetic Interval (OGI) to Sequence Assembly
– Ontology Applying to Next Generation Sequencing

Yu Lin[1], Hiroshi Tarui[1], Peter Simons[2]

1. Genome Resource and Analysis Unit, Genomics Laboratory, Center for Development Biology, RIKEN, Japan {linyu,tarui}@cdb.riken.jp
2. Department of Philosophy, Trinity College Dublin, Ireland psimons@tcd.ie

Abstract. We develop an OWL ontology: OGI (Ontology for Genetic Interval) for the formalization of the genomic elements by defining them as a Genetic Interval. Based on OGI's definition of Genetic Interval Relations, which derived from the Allen interval calculus, we attempt to represent the relationships among contigs and sequence data from next generation sequencing. A real dataset generated from the bench has been loaded and tested by using SPARQL for validating OGI. Although the dataset is small, this semantic-based method provides a clue for assembly sequence. Evaluating this method on a bigger dataset, both harmony and conflict of definitions with current ontologies, such as SO (Sequence Ontology), need to be considered. OGI is available on NCBO's BioPortal website.

Keywords: OGI, ontology, genetic interval relations, next generation sequencing, 454 FLX sequencer, SuperContig

1 Background

The mother ontology for genetic interval is Ontology of Genetic Susceptibility Factors (OGSF), which has been built as a modular ontology of the ontologies for genetic susceptibility to disease [1]. OGSF includes three ontologies: Ontology of Genetic Susceptibility Factors (OGSF), Ontology of Glucose Metabolism Disorder (OGMD) and Ontology of Geographical Region (OGR). When we discovered that the co-localization of genetic susceptibility factors (such as SNPs) with the gene is the major criterion for determining a susceptibility gene, we developed another ontology: Ontology for Genetic Interval (OGI, http://bioportal.bioontology.org/ontologies/40117). The purpose of OGI is to formalize the genomic elements, including genes, mutations mRNAs, and all kinds of genomic structures which are essential to current genomics research.

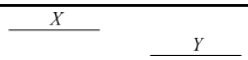
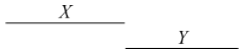
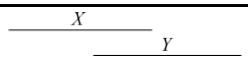
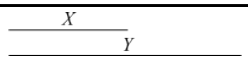
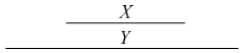
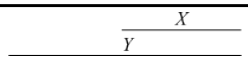
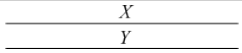
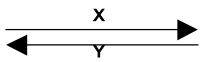
2 Introduction of OGI

In current molecular biology research, the gene and other genomic elements have been modeled as a sequence, which is represented as the combination of A, T, C and G in a linear form. Ontology for Genetic Interval described the Genetic Interval as a subclass of Biological Interval, which is a "spatial continuous physical entity which contains ordered biological sets (DNA segment, Nucleic Acid Base Residue, RNA segment, Protein segment) between two boundaries: start boundary and end boundary on a chromosome, RNA or protein".[2] The difference between Genetic Interval and Genetic Sequence is that Genetic Interval holds not only the primary linear sequence information but also the 3-

D structure information of a given Genetic Interval. However, since the current capability of genomics research is limited to the linear information of a Genetic Interval, the genetic sequence is taken into account as a simplified model of genetic interval.

In OGI, we define a sequence as a specific kind of group (collective) within the methodological and ontological constraints of nominalism. In order to conceptualize the notion of sequence we start from logic, which is an indispensable part of constructing an ontology. Then we define the Genetic Interval Relations by borrowing ideas from the Allen interval calculus (Table1).

Table 1. Genetic Interval Relations

Relations in Allen Interval	Illustration	Relations of Genetic Interval
X<Y Y>X		isLocatedBefore (xLBy) isLocatedAfter (yLAx)
XmY YmiX		isAdjacentBefore (xABy) isAdjacentAfter (yAAx)
XoY YoiX		isOverlapStartWith (xOSy) isOverlapEndWith (yOEx)
XsY YsiX		isStartsWith (xSWy,ySWx) (symmetric property)
XdY YdiX		isContainedIn (xCIy) (transitive property)
XfY YfiX		isEndWith (xEWy, yEWx) (symmetric property)
X=Y		isEqualTo (xEy, yEx) (symmetric property)
		isReverseCompleteOf (xRCy) (symmetric property)

3 Next Generation Sequencing and Sequence Assembly

According to wikipedia, "The term DNA sequencing refers to sequencing methods for determining the order of the nucleotide bases—adenine, guanine, cytosine, and thymine—in a molecule of DNA." [3] State-of-the-art next-generation sequencing platforms such as the Roche-454 GS FLX, Illumina Genome Analyzer and ABI SOLiD provide high-throughput and high-speed technology to read the nucleotide bases of samples. However, the reading length generated by such sequencers is very short: ~400bp by Roche-454 FLX titanium, ~75bp by Illumina, and ~50bp by ABI SOLiD. The previous and widely used sequencer is Sanger 3730 series, by which a reading length up to 2000bp can be obtained.

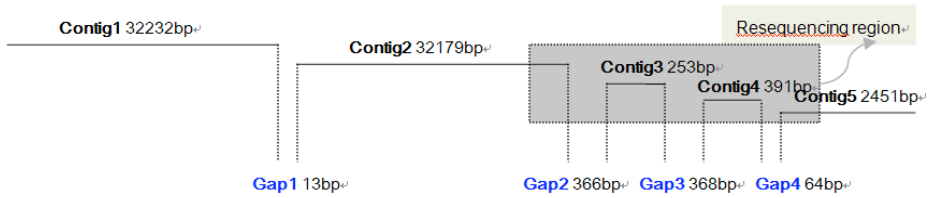
Usually, a sequencing project is taken by randomly cutting the target sequence into smaller fragments, which is the so-called "shotgun method"; and then after next generation sequencers obtain the readings, a computer will attempt to put all readings together to give the whole view of the target sequence. Thus, the assembly procedure for combining the readings either by mapping or de novo method is essential. How to get a closest overall picture of the target sequence is a key issue for both bioinformaticians and biotechnologists.

4 A Practice of Using OGI to Assemble the Readings

4.1 Method

In this experiment, the length of target sequence is up to 60k bp. The shotgun sequencing method was applied to get the fragments of sequences; after amplification by PCR, all the fragments were then mixedshothe up for building a DNA library for running on a 454 FLX sequencer. After we got the raw readings dataset from the 454 FLX sequencer, we used the software package provided by 454 Life Sciences Corporation. Since the sample comes from a model organism, we chose the GS Reference Mapper Application from the package to assemble reading. All the ~300bp readings were assembled as 5 contigs. (Contig: a group of overlapping readings derived from a single genetic source.) Comparing the contigs with the reference genomic sequence by using a similarity alignment method, we found out that there are four gaps between those contigs. According to the length of the Gaps and Contigs, a resequencing region flanking the ending of Contig2, Gap2, Contig3, Gap3, Contig4, Gap4 and the beginning of Contig5 has been decided. (Fig. 1.)

Fig. 1. Contigs and Gaps



Except for Gap1, which can be filled by manually checking the sequence, we resequenced the resequencing region above by the shotgun method using a Sanger 3730 sequencer. After repeating the alignment and reference sequence comparison steps as in the previous stage, relationships between those Contigs and Gaps were represented using Genetic Interval Relations from OGI. Then, all these entities (Contig and Gap) and their relations were manually populated into OGI under the software environment of Protégé 4.0.

4.2 Result

4.2.1 Representing the relations between Contigs and Gaps

Contig1 isLocatedBefore Contig2
Contig2 isLocatedBefore Contig3
Contig3 isLocatedBefre Contig4
Contig4 isLocatedBefre Contig5
Gap2 isAdjacentBefore Contig3
Gap3 isAdjacentBefore Contig4
Gap4 isAdjacentBefore Contig5
Contig1 isOverlapStartWith Contig2
Contig2 isOverlapSartWith RevReseqD12
Contig3 isContainedIn RevReseqD12
Gap2 isContainedIn RevReseqD12
RevReseqD12 isOverlapEndWith Gap3
Contig4 isOverlapSartWith ReseqE12
Gap4 isContainedIn ReseqE12
ReseqE12 isOverlapEndWith Contig5
Contig3 isOverlapStartWith RevReseqF12
Gap3 isContainedIn RevReseqF12

Contig4 isContainedIn RevReseqF12
Gap4 isContainedIn RevReseqF12
RevReseqF12 isOverlapStartWith Contig5
RevReseqF12 isReverseCompleteOf ReseqF12
RevReseqD12 isReverseCompleteOf ReseqD12

4.2.2 Axioms which are built in Protégé 4.0 as follows:

```
Axiom1: if Z isOverlapStartWith X and X isContainedIn Y,
then,
Z isOverlapEndWith Y => Y isOverlapStartWith Z
```

```
Axiom2: if Y isOverlapEndWith X and X isContainedIn Z, then,
Y isOverlapStartWith Z => Z isOverlapEndWith Y
```

Using the pellet reasoner from Protégé 4.0, the following relations are inferred by the software:

RevReseqD12 isOverlapStartWith RevReseqF12
RevReseqF12 isOverlapStartWith ReseqE12

4.2.3 SuperContig analysis

A SuperContig means complete nucleotide sequence information of the sample, in our case, the 60kbp's target sequence. The SuperContig should start with Contig1 and end with Contig5, which also means that we are looking for a path which starts from Contig1 and ends with Contig5.

SuperContig
    isStartWith Contig1
  isEndWith Contig5

SuperContig is an accumulation of all the assembled Contigs, and all the subsets relations in this SuperContig must be the same as each other: isOverlapStartWith .

Fig. 2. Path for building a SuperContig

(The red dashed line shows the path to generate a SuperContig in Fig.2).

- Using an open source reasoning tool kit LSW[4], which uses Pellet as the underlying reasoner, we ran SPARQL codes and got following results:

```
SELECT ?x ?y ?z
WHERE {
        ?x OGI:isOverlapStartWith OGI:Contig5 .
        ?y OGI:isOverlapStartWith ?x .
        ?z OGI:isOverlapStartWith ?y .
        OGI:Contig1 OGI:isOverlapStartWith ?z . }
```

```
Results:  OGI:RevReseqF12 OGI:RevReseqD12 OGI:Contig2
```

- Thus, the SuperContig were constructed by (in the correct order):

```
Contig1 isOverlapStartWith Contig2
Contig2 isOverlapSartWith RevReseqD12
RevReseqD12 isOverlapStartWith RevReseqF12
RevReseqF12 isOverlapStartWith Contig5
```

∴ SuperContig will be generated by overlapping the nucleotide sequence of Contig1, Contig2, RevReseqD12, RevReseqF12, and Contig5.

5 Conclusion

In this study, we have populated a small dataset of contigs assembled from the 454 FLX sequencing readings to OGI; the purpose is to construct a SuperContig which can give the complete nucleotide sequence information of the target sequence. Pellet reasoner, logic rules and SPARQL language were applied for finding the path for building SuperContig. Being able to form one SuperContig out of all the readings is an essential step in sequencing method. Many assembly tools are using either mathematical or object-oriented methods to construct the SuperContig. Here we tried to follow the semantic method, which helps people understand the relations of contigs, especially those resequencing contigs or fragments generated by different sequencers, in our case, both next and first generation sequencer used for a deep and accurate sequencing.
Since our method applied to the limited contigs generated by standard software rather than the huge readings, it is practical and reasonable. However, when dealing with a whole genome, especially those which have no reference genome sequences to compare with, more contigs will complicate the assembly. A better solution will be needed for accelerating the capability of reasoning and performance.

Another issue is that the overlap between OGI and current Sequence Ontology (SO) [5] is unavoidable. Although not yet discussed with the consortium of Sequence Ontology, OGI contributes to SO by providing the relations described in this paper. Authors of OGI are planning to merge SO into OGI by adopting the terms from SO, and it will be important to report a detailed mapping between those two ontologies.

References

1. Lin Y., Sakamoto N.: Ontology of Genetic Susceptibility Factors to Diabetes Mellitus (OGSF-DM). Interdisciplinary Ontology Proceedings of the First Interdisciplinary Ontology Meeting. 99 --104 (2008)
2. Lin Y., Sakamoto N.: Genome,Gene,Interval and Ontology. Interdisciplinary Ontology Proceedings of the Second Interdisciplinary Ontology Meeting. 25—34 (2009)
3. http://en.wikipedia.org/wiki/DNA_sequencing
4. http://esw.w3.org/topic/LSW LSW is an open source set of lisp tools for working with OWL and SPARQL using the Pellet reasoner. It was initially written by Alan Ruttenberg, but is starting to accumulate contributions from others.
5. http://www.sequenceontology.org/ The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence.