Clonal Complexes in Biomedical Ontologies

Albert Goldfain[1], Lindsay G. Cowell[2], Barry Smith[3]
[1]Blue Highway, Syracuse, NY; [2]Duke University Medical Center, Durham, NC; [3]University at Buffalo, Buffalo, NY

Abstract

An accurate classification of bacteria is essential for the proper identification of patient infections and subsequent treatment decisions. Multi-Locus Sequence Typing (MLST) is a genetic technique for bacterial classification. MLST classifications are used to cluster bacteria into clonal complexes. Importantly, clonal complexes can serve as a biological species concept for bacteria, facilitating an otherwise difficult taxonomic classification. In this paper, we argue for the inclusion of terms relating to clonal complexes in biomedical ontologies.

Introduction

Many of the difficulties in classifying bacteria stem from the fact that bacteria are both biological organisms (subject to biological classification) and, in certain circumstances, pathogens (subject to a disease-based classification). The fact that a bacterium can play the role of pathogen is an important ontological fact, but entities should not be classified solely on the basis of the roles they can play. If there is to be a bias in classifying bacteria, it should be a biological bias. This provides a more uniform classification scheme for all biological organisms.

Adopting such a classification brings up the problem of how to treat species at the microbiological scale. Following Mayr[1], we adopt the biological species concept in which differing species are separated by a barrier to gene flow.

Multi-Locus Sequence Typing (MLST) and Biomedical Ontologies

MLST is a popular method for achieving a biological classification for bacteria by using the allelic differences of seven housekeeping genes to determine the degree of relatedness between strains. "Clonal expansion [for bacteria] results from the rise in frequency of a single highly adaptive genotype. These ancestral genotypes subsequently diversify through recombination or mutation to produce minor clonal variants, and hence a 'complex' of closely related strains.[2]" The BURST clustering algorithm uses MLST data to infer this ancestral genotype and assign observed genotypes to clonal complexes. Seven genetic housekeeping loci are selected for a given genotype pool and the ancestral genotype is defined to be "the genotype within the clonal complex that differs from the highest number of other genotypes in the clonal complex at only one locus out of seven [these are called single locus variants (SLV)].[2]" The BURST algorithm succeeds when the assigned ancestral genotypes match the actual ancestral genotype in the phylogeny of the bacteria. BURST output is usable as biological species demarcation when the complexes are genetically isolated as illustrated in Figure 1.

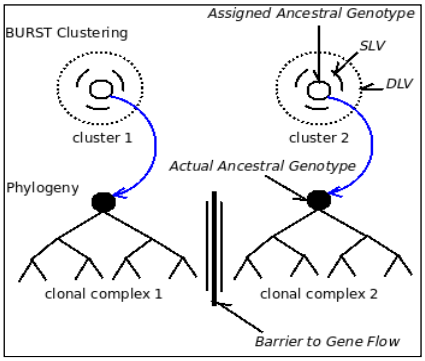

Figure 1. BURST clustering into clonal complexes

This technique does not always produce crisp demarcations between clonal complexes due to horizontal gene transfer. However, any biological taxonomy must tolerate some vagueness and fuzzy borders.

The inclusion of clonal complexes in biomedical ontologies requires the inclusion of several other terms: *clone*, *isolate*, *strain*, *housekeeping gene, ancestral genotype*, *recombination*, and *clonal divergence*. The adoption of these terms will yield a more uniform treatment of biotic entities of all sizes and will furnish a sound biological basis for disease ontologies.

Acknowledgements

This work was funded by the National Institutes of Health through Grant R01 AI 77706-01. Smith's contributions were also funded through the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028 (National Center for Biomedical Ontology).

References

1. Mayr E. The Species Category. Toward a New Philosophy of Biology. 1988; 315—334.
2. Feil E and Man-Suen C. The BURST algorithm. http://pubmlst.org/analysis/burst/burst.shtml