A natural programming language to study repeated DNA sequences

Jeff Elhai
Center for the Study of Biological Complexity

In the last 10 years, biologists have been bestowed with an astonishing amount of information, an amount that demands automated methods to comprehend. Unfortunately, they have been slow to devise methods suitable for their particular problems because few biologists program computers. Instead, those that make use of mass information generally rely on a few tools that channel their perceptions along standard lines. It is not practical to ask biologists to learn general computer languages. So, if the biologist will not come to the language, the language must come to the biologist.

BioLingua is a new general-purpose computer language that has a syntax particularized to the needs of the molecular biologist. It draws on a database (ideally) comprehending all available information of interest to a specific discipline and provides a standard interface to access that information. More important, it provides a language that permits molecular biologists to manipulate that knowledge in ways that are verbally similar to thought processes they are used to. The presentation will stress the repeating cycle of low-level analysis (Fig. 1) and high-level visualization (Fig. 2) that is necessary for human intuition to connect productively with reality.

The difficulties of gathering and manipulating information from varied sources and the contrasting joy of doing the same tasks within BioLingua will be illustrated with an analysis of repeated sequences from the cyanobacterium *Nostoc punctiforme*. Repeated sequences comprise much of the poorly understood intergenic regions (97% of the human genome) that is responsible for genetic regulation and genome plasticity. *Nostoc*, like higher eukaryotes but unlike other bacteria, possesses a bewildering collection of tandem repeat families, as well as hundreds of copies of short dispersed sequence units. The large number of identical repeats force upon us the questions: Where did they come from? How do they propagate themselves? and What functions do they serve?

```
(LOOP FOR protein IN (PROTEINS-OF Npun)
        AS mw = (MW-OF protein)
        AS ratio = (RATIO-OF protein
                      FROM Hihara2001 COLUMN 1)
      WHEN (AND (EXISTS ratio)
                (> mw 48000)
                (< mw 52000)
                (> ratio 2))
      COLLECT (LIST protein mw ratio))
```

Fig. 1: Example of BioLingua code. Suppose you found a provocative spot on a gel of *Nostoc* proteins whose intensity varies with experimental conditions. What might that protein be? To answer that question, you consider each protein of *Nostoc*, calculate its molecular weight, look up its expression in a microarray experiment that addressed the experimental condition, and define a set of all proteins that have the desired molecular weight and desired degree of induction. The result is a list of candidate protein and their properties. The underlined words are BioLingua functions. Within BioLingua, clicking on such links brings the user to documentation of their function

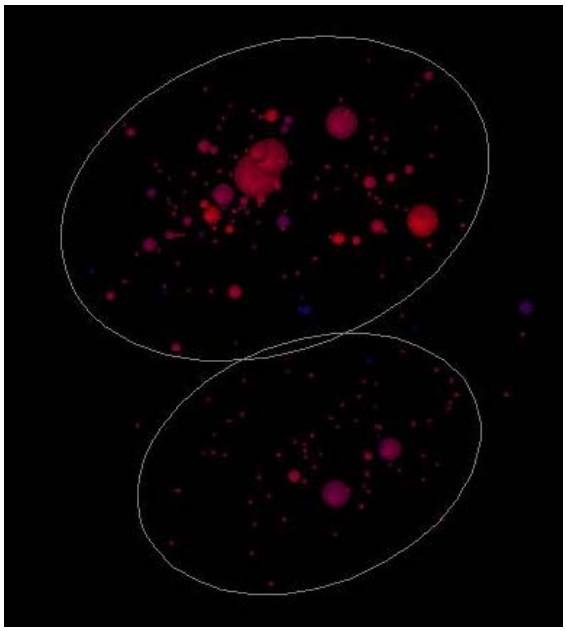

Fig. 2: Visualization of relatedness between members of repeated sequence NIS-1 family. Size of ball represents copy number. Distance between pairs represents number of mismatches.