1 NCBI: INTEGRATED DATA FOR MOLECULAR BIOLOGY RESEARCH

Karl Sirotkin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Summary

Since 1992 the National Center for Biotechnology Information (NCBI) has provided integrated access to all public genetic sequence information and its associated annotation, as well as the citations and abstracts from the published literature referenced by the genetic sequence records. This chapter describes the main database that contains the genetic sequence data used by this integrated access, and how these data are linked both to other sequences and to the published literature. Entrez is the application used for accessing most of these data and their links and it can be used on a wide variety of hardware platforms. More recently, Web browser-based Entrez access has also been available (URL: http://www.ncbi.nlm.nih.gov/Entrez/). Subsets of these data are also available for Blast searching. It is hoped that this chapter will be a useful resource for both software developers and end users of this data.

Introduction

Interconnections between data help to integrate them. NCBI defines interconnections between genetic sequence data, structure data, taxonomic data, and literature references. These links may also be between the same type of records, for example, between literature articles. Articles are linked as similar using term content statistics [1,2,3]. Links between genetic sequence records are based on Blast sequence comparisons [4], linking similar, and thus possibly homologous sequences. Links between structure records are based on Vast structure comparisons [5], linking structures that tend to be similar.
Each record of one type can also link, reciprocally, to records of other types. Citations in the genetic sequence records determine links between the literature and genetic sequence information. Proteins translated from nucleic acid sequences, are linked to those sequences. Sequences can also be retrieved based on the phylogeny of the organism from which they come.

The data integrated by NCBI has recently been expanded. Now, most of the biological literature is available in a new database, PubMed, that includes all MEDLINE records plus some additional records. Links to electronic full-text articles have also been added where possible. Genomic and chromosomal maps linked to sequence information are available. Curated taxonomic tree is maintained [6] for organisms identified by the authors of the sequence submissions as the source of the genetic sequence data. This tree is independently available on the web (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomy/home.html). Subsets of sequence records from organisms at taxonomic subtrees can be retrieved.

The rest of this chapter describes the management of the sequence data interconnected by NCBI. This data is represented by the two lower hexagons in

---

**Figure 1**: Data interconnected by NCBI.

*Figure 1: Hexagons represent data that has been integrated, curved lines represent links between similar records, and straight lines between hexagons represent between different data types, reciprocally linked, as described in the text. Dotted lines represent links under development.*