Bioinformatics Research at CSCL, IIS, Academia Sinica

Dr Jan-Ming Ho (PhD) and Dr Wen-Dar Lin (PhD)

Computer Systems and Communication Lab (CSCL), Institute of Information Science (IIS), Academia Sinica, Taiwan, has dedicated to the fields of web mining, communications, and media technologies for several years. Started three years ago, members of CSCL, Dr Jan-Ming Ho, Dr Wen-Dar Lin, and their colleagues, have also put their efforts in bioinformatics research through collaborations with biologists in Academia Sinica. Their previous experiences include (1) working with Genotyping Center of Institute of Biomedical Science (IBMS) and Computing Center of Academia Sinica to design and implement a laboratory information management system (LIMS) prototype; (2) working with Dr Wen-Chang Lin’s team at IBMS to design and implement an EST annotation system, Bio101, which provides friendly user interfaces for biologists to manage and analyze their EST data; (3) working with Dr Ming-Jing Hwang’s team at IBMS to develop the UniMarker based synteny mapping technology.

Recently, the group’s research interests include three main directions: (1) index-based genome alignment methods, (2) biological database integration tools, and (3) computational prediction of biological elements. Each direction includes one or more ongoing projects, and we list some individual projects are listed as follows:

(1) An Efficient Synteny Mapping and Annotation System

Synteny mapping, or detecting regions that are orthologous between genomes, is a key step in comparative genomics. The goal is to develop an efficient synteny mapping and annotation system for vertebrate genomes based on the UniMarker-synteny experience.

The group has developed an efficient and effective synteny mapping method on two large genomes by using highly conserved uni-makers, i.e., fixed-length DNA fragments that appear only once in each genome, as indices. The occurrence spectra of these 16-mer uni-makers in human and mouse genome are then used to detect orthologous genomic segments. It avoids the need to spent large amount of time in computing genome alignment. The method indeed runs very fast, e.g., mapping mammalian genomes of human and mouse takes one day of computing time on a single Pentium IV personal computer. The resulting human–mouse synteny map was shown to be in excellent agreement with those produced by the Mouse Genome Sequencing Consortium (MGSC) and by the Ensembl
team; furthermore, the syntenic relationship of segments found only by
their method was also supported by BLASTZ sequence alignment results.
This work is also published as a public open-source project. They are
continuing to explore a new family of markers that can be used to detect
synteny maps of distant species in which the density of uni-markers is
low.

(2) EST Annotation Pipeline System

EST sequencing provides a means for studying functional genomics
under limited resources. The primary goal of an EST annotation pipeline is
to annotate the functions of a query EST by aligning it with existing gene
databases of known species. However, to compensate for errors associated
with the query ESTs, a complete EST annotation pipeline consists of three
components: sequence cleaning phase, sequence clustering/assembling
phase and annotation phase. Bio301, after its predecessor Bio101 is the
second EST annotation pipeline developed at CSCL. Bio301 annotates
ESTs in standard terms defined in Gene Ontology. Bio301 also provides
biologists with post-processing tools to further analyze, data mining and
visualize the annotation results, e.g., digital differential display (DDD)
analysis, some statistical analysis, and GOBU. GOBU is a visualization
tool for biologists to systematically and simultaneously browse the
annotation results along several dimensions of EST properties. More details
on GOBU are given in the next section.

(3) Gene Ontology Browsing Utility

Since the inauguration of Gene Ontology (GO) consortium, the
number of gene products annotated in GO terms has been increasing. In
order to facilitate biologists to browse and to manipulate GO-annotated
gene products systematically, they designed a Java-based software called
Gene Ontology Browsing Utility (GOBU). The design of GOBU is based
on the observation that the GO terms for describing gene functions are
organized in a hierarchical structure, or more precisely as a directed
acyclic graph. This is also true for some other properties that are associated
with gene products by other bioinformatics tools. That is, the vocabulary
for describing such a property can be specified in a tree structure. Thus
they designed GOBU such that it accepts a set of gene products and the
associated tree-structured vocabulary of each property of interest. Users
then can browse and manipulate the given gene products in multiple
interactive windows each being associated with a given property.

(4) Matching Protein domains with EST

Most of the previous annotation pipelines annotate an EST sequence
by searching for homolog or highly similar sequence in an existing gene
database. In this study, the group has developed a new EST annotation
tool, denoted as E2D, by searching for a protein domain that best matches
given EST sequence. Note that a protein domain usually exists in several
protein sequences. They detect the association of a protein domain to the
EST by taking advantage of coherence of sequence alignment between
these protein sequences and the EST sequence.
Our experiments show that E2D gives satisfactory annotation results and is fast compared to tools with similar goals, e.g., InterProScan. Another side benefit is that E2D does not require the query EST to contain a complete domain site.

(5) Predicting protein 3D structure from 1D sequences

Three-dimensional (3D) structure of a protein is a critical determinant of its biological function. Predicting 3D structures from one-dimensional amino acid sequences is a core challenge in biology. A number of different approaches have been developed to address this problem, but considerable challenges remain. In recent years, some success has been achieved using a knowledge-based approach in which an intermediate component (local structures or fragments) is introduced. It splits the problem into two parts: a) mapping fragments in terms of sequence-to-structure relationship and b) assembling fragments. The first problem usually results in the prediction of multiple but few fragments for a specific sequence; subsequently, a combinatorial search among these fragments is used to make a prediction of an entire protein structure by piecing together native-like protein structures according to energy function constraints.

The research group had established a protein fragment library as the basis of structural building blocks and are working on the prediction problem.

Web links

1. The UniMarker-Synteny method: http://syntenymap.openfoundry.org/ or http://synteny.iis.sinica.edu.tw/UM
2. Bio301 EST annotation pipeline: http://bc02.iis.sinica.edu.tw/bio301/. This is a prototype website of our next generation EST annotation system. Users may use the account “demo” to login.
3. GOBU GO browsing utility: http://bc02.iis.sinica.edu.tw/, a Java program for generic GO browsing and manipulation.
4. E2D EST domain recognition tool: http://bc03.iis.sinica.edu.tw/LEE/E2D.html