A Survey of Bioinformatics

Humberto Ortiz Zuazaga
University of Texas Health Science Center at San Antonio
humberto@momo.uthscsa.edu
http://momo.uthscsa.edu/~humberto/cmb/

May 8, 1997

Bioinformatics

The application of computers to the collection, analysis, and presentation of biological information.

Taxonomy (of the talk)

- Sequence analysis
- Structure prediction
- Genetic and physical mapping
- Gene expression networks
- Biological databases
- Biological software components

Sequence analysis

- Pairwise sequence alignment
- Multiple sequence alignment
- Phylogenetic analysis
- Gene prediction

Pairwise sequence alignments

Given a pair of sequences over an alphabet Σ, and a cost function that assigns a cost to an alignment, find an alignment of the strings that minimizes the cost.

- This problem is central to many biological programs (homology searches, multiple alignments, molecular phylogeny, exon prediction, protein threading, ...)

- Usually solved by Dynamic Programming.

Multiple sequence alignments

- Assign a cost to alignments of multiple sequences.

- Sum of pairs metric, add all pairwise alignments together.

- NP-complete for the SP metric.

- Other heuristics must be used for practical programs (Divide and Conquer, HMM's).

Structure prediction

- Theoretical models of protein folding
- Statistical models of protein folding

The H-P model of globular proteins

- Amino acids are hydrophobic (H, nonpolar) or hydrophilic (P, polar).

- H-H contacts contribute -1 to the energy of the protein.

- all other contacts contribute 0.

- Protein structures are constrained to self-avoiding paths on a regular lattice.

The Inverse Protein Folding (IFP) problem

Given a target structure or conformation of a protein G, find a sequence s of length n that:

- Has G as is it's minimum energy state.
- Has the lowest degeneracy (number of other conformations with the same energy) of any possible sequence.

The Heuristic Sequence Design (HSD) problem

IFP is conjectured to be NP-complete, the best known algorithm must search over all possible conformations of all possible sequences. HSD problems try to simplify the computation by restricting the problem.

- The Canonical Method: find the sequence with at most λn hydrophobic residues.

- The Grand Canonical Method: Change the energy function so that H-H contacts have -2, an exposed H residue has 1, and all other interactions have 0.

Computational Results

- The Canonical Method is NP-complete (reduction from SUBSET-SUM), but there are algorithms that can approximate the energy of the optimal solution (OPT) to 1+OPT on 2D lattices, and 1/2 OPT on 3D lattices.

- The Grand Canonical Method can be solved in polynomial time.

Protein Threading

- Divide each sequence into functional domains (by multiple alignments).

- Thread each domain into a structure from a library of similar domains.

- Compute a "pseudo-energy" based on the empirically determined probabilities that each structural unit in the protein would be in the structural environment it is in.

Genetic and physical mapping

- Maximum likelihood mapping

- Breakpoint mapping

- Sequencing by hybridization

Gene expression networks

- Complete genomes available for several species.

- 100,000 human genes, many already sequenced.

- Large scale synthesis of thousands of oligo probes.

- Biologists will soon be able to measure expression levels for ALL GENES in a single assay.

Biological databases

- ACEDB

- ECO-CYC

Biological software components

- bioTk
- bioWidgets for Java
- bioViews