Bioinformatics Tools

Byoung-Tak Zhang and Chul Joo Kang
School of Computer Science and Engineering
Seoul National University

Contents

1. Sequence Alignment
2. Multiple Sequence Alignment
3. Pattern Finding
4. Structure Prediction
5. DNA Microarray
6. Major Tools for Proteomics
Bioinformatics Tools

<table>
<thead>
<tr>
<th>Problems</th>
<th>Tools or Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence Alignment</td>
<td>BLAST, FASTA</td>
</tr>
<tr>
<td>Multiple Sequence Alignment</td>
<td>Clustal W, Macaw</td>
</tr>
<tr>
<td>Pattern Finding</td>
<td>GRAIL, FGENEH, tRNAscan-SE, NNPP, eMOTIF, PROSITE, ChloroP</td>
</tr>
<tr>
<td>Structure Prediction</td>
<td>Bend.it, RNA Draw, NNPREDICT, SWISS-MODEL</td>
</tr>
<tr>
<td>DNA Microarray</td>
<td>GeneX, GOE, MAT, GeNet</td>
</tr>
<tr>
<td>Hardwares for Proteomics</td>
<td>2D Gel, MALDI-TOF</td>
</tr>
</tbody>
</table>

Bioinformatics Based on Sequence

- Genome sequencing project: get huge sequence data
- Retrieving useful information from sequence data
  - Find genes and other elements
  - Classification
  - Predict its function
  - Others
- Read the articles on Nature and Science
1. Sequence Alignment

Pairwise Sequence Alignment

- Basic job handling sequences
- Alignment between two sequences
- Sequence database search
- Finding similar DNA or protein sequences

- Smith-Waterman algorithm
- BLAST
- FASTA

Score = \sum_{Region Start}^{Region End} Similarity weights - \sum_{Region Start}^{Region End} Penalties
Smith-Waterman Algorithm (1)

- Finding local alignments
- Using dynamic programming

\[
\begin{align*}
S(i-1,j-1) + s(i,j) & \\
S(i-1,j) - GP & \\
S(i,j-1) - GP & \\
V(i-2,j) - GEP & \\
H(i,j-2) - GEP & \\
0 & \\
\end{align*}
\]

\[s(i,j) = \text{Max}\]

TCAT*G
*CATTG

Smith-Waterman Algorithm (2)

- Best results and slow performance
- Can grasp results missed by BLAST or FASTA.
- Available on web: spiral.genes.nig.ac.jp/homolgy/ssearch-e.shtml
BLAST & FASTA

- Using heuristic algorithms
- Word based match
- Faster than Smith & Waterman
- **BLAST**: www.ncbi.nlm.nih.gov/BLAST
- **FASTA**: www.ebi.ac.uk/fasta3

---

The BLAST Search

For the query, find the list of high scoring words of length W.

For each word from the query sequence, find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).

For each word match, extends the alignment in both directions to find alignments that score greater than a threshold of value S.

Exact matches of words from word list

Database Sequences

Word List

Maximal Segment Pairs (MSPs)
BLAST Result

Program: BLASTP
Query: human MYB binding protein
Database: Swissprot

2. Multiple Sequence Alignment
Multiple Sequence Alignment

- Align multiple sequences
- Finding well conserved regions
  - Finding motifs or other elements
- Building gene families
- Constructing a phylogenetic tree
- Using whole dynamic programming?
  - \( O(n^k) \) problem (\( n = \) sequence length, \( k = \) number of sequences)
- Clustal W
- Macaw

Clustal W

- Step 1: Pairwise alignment between two sequences
- Step 2: Making sequence weights and clusters
- Step 3: Alignment between most similar sequences or clusters
- Step 4: Making one optimal output sequence

Multiple Alignment by Clustal W

Phylogenetic Tree Construction Based on Clustal W Results
Pattern Finding

- DNA sequences have all information about their products
  - Gene finding
  - Other sequence element finding
  - Motif finding
  - Localization prediction
  - And others
Gene Structure

- From an unannotated DNA sequence, find putative expressive regions
- Distinguish exon and intron regions (in eukaryote).
- Solutions
  - Compare known genes (alignments)
  - Simply find conserved region
  - Using statistical models like hidden Markov models and others
GRAIL: Find Exon and Intron

- In eukaryotes, RNA is processed.
- Intron: splice out
- Exon: join together
- GU – AG rule
- GRAIL: exon prediction from a genomic sequence with RepeatMasker filtering
- Neural networks which combine a series of coding prediction algorithms
- http://compbio.ornl.gov/Grail-1.3

GRAIL Results
FGENEH: Prediction of Multiple Genes in Human DNA Sequences

- Finding genes from DNA sequences
- HMM based human gene structure prediction
- Available on web:
  http://searchlauncher.bcm.tmc.edu/seq-search/gene-search.html
tRNAscan-SE: Find tRNA Genes

- Finding tRNA from genomic sequence
- Finding polII promoter sites, searching conserved secondary structures - easier problem than finding protein coding genes
- Combining several earlier programs and algorithms
- Http://www.genetics.wustl.edu/eddy/tRNAscan-SE/
Sequence Element Finding

* Promoter region of E.coli lacZ operon

- Promoter and other elements: regulatory and other protein binding sites

NNPP: Neural Network Promoter Prediction

- Finding promoter sequences on DNA sequences
- Time-delay neural networks
- Two feature layers
  - Recognizing TATA boxes
  - Recognizing initiators
- [http://www.fruitfly.org/seq_tools/promoter.html](http://www.fruitfly.org/seq_tools/promoter.html)
Neural Network Promoter Prediction

Motif Finding

- Find conserved residues from DNA or protein sequences.
- Commonly, conserved residues are related to protein functions.

Solutions:
- Finding consensus sequences
- Alignment
- Position specific weight matrix
- Hidden Markov models

Motif of one class of zinc finger protein

\[(\text{C}_2\text{H}_2) \quad \text{Cx}\{2,4\}\text{Cx}\{12\}\text{Hx}\{3,5\}\text{H}\]
Motif Databases: eMOTIF, PROSITE

- Prosites, eMotif: protein motif databases
- eMotif (motif.stanford.edu/emotif)
  - eMotif Maker: building new motif from multiple alignment
  - eMotif scan: find motif pattern
  - eMotif search: find motif from protein sequence

eMOTIF (2)

In the Swiss-Prot database:

**ZEP_HUMAN**
ZINC FINGER PROTEIN 4D (HUMAN IMMUNODEFICIENCY VIRUS TYPE I INTEGRASE-BINDING PROTEIN 1) (HIV-1) (MAJOR HISTOCOMPATIBILITY CLASS I POLYMERASE)

**ZEP_MOUSE**
ZINC FINGER PROTEIN 4D (TRANSCRIPTION FACTOR) (ALPHA-CRYP1) (ALPHA CRYP1)

**UTL_HUMAN**
KINEASE-LIKE PROTEIN KIFIA (KINESIN TRANSPORTER OF SYNAPTIC VESICLES)

**ULS_MOUSE**
ZINC FINGER PROTEIN BL13

**ULS_HUMAN**
ZINC FINGER PROTEIN BL13

eMOTIF (3)

At an expectation of 0.01 or less:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Expectation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2.38e-12</td>
<td>Zinc finger, C2H2 type c.k.p.p.r.l.a.[-g][d][k.p.p][g.c] c.38-C00978999THNIFHIFVFA-O-63</td>
</tr>
<tr>
<td>2.</td>
<td>3.00e-11</td>
<td>Zinc finger, C2H2 type c.k.p.p.r.l.a.[-g][d][k.p.p][g.c] c.38-C00978999THNIFHIFVFA-O-63</td>
</tr>
<tr>
<td>3.</td>
<td>3.00e-11</td>
<td>Zinc finger, C2H2 type c.k.p.p.r.l.a.[-g][d][k.p.p][g.c] c.38-C00978999THNIFHIFVFA-O-63</td>
</tr>
<tr>
<td>4.</td>
<td>3.95e-10</td>
<td>Zinc finger, C2H2 type c.k.p.p.r.l.a.[-g][d][k.p.p][g.c] c.38-C00978999THNIFHIFVFA-O-63</td>
</tr>
<tr>
<td>5.</td>
<td>3.95e-10</td>
<td>Zinc finger, C2H2 type c.k.p.p.r.l.a.[-g][d][k.p.p][g.c] c.38-C00978999THNIFHIFVFA-O-63</td>
</tr>
<tr>
<td>6.</td>
<td>3.95e-09</td>
<td>Zinc finger, C2H2 type c.k.p.p.r.l.a.[-g][d][k.p.p][g.c] c.38-C00978999THNIFHIFVFA-O-63</td>
</tr>
</tbody>
</table>
PROSITE (1)

- PROSITE: Database of protein families and domains (http://www.expasy.ch/prosite/)
- Consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

PROSITE (2)

e.g.) C2H2 zinc finger protein
ScanPROSITE

- Compares query sequences (protein) to Prosite.
- Find patterns

Others: Find Gene Destination

- After translation, proteins are located to proper places (nuclear, mitochondria, outside of cell...)
- That process determined by amino acid sequences on proteins
  - Finding signal peptide
  - Predicting protein targets
ChloroP: Prediction of Chloroplast Importing Protein (1)

- Predicting chloroplast targeting transit peptides
- Predicting cleavage sites
- Neural network based method

- Service site: 
  http://www.cbs.dtu.dk/services/ChloroP/

(C) 2001 SNU CSE Artificial Intelligence Lab (SCAI) 39

ChloroP: Prediction of Chloroplast Importing Protein (2)
4. Structure Prediction

Structure Prediction

- DNA, RNA and protein structure prediction
- These structures affect their function, stability, etc.
- Protein structure prediction by calculation of biochemical properties of each amino acid
  - In most cases of protein secondary or tertiary structure prediction, it is a terribly huge computing job (almost impossible)
- Prediction based on known structures
DNA & RNA Structure

- DNA & RNA strands: dynamic one, not linear form
- DNA: regionally denatured (breathing), bending
  - Affects its expressions and others.
- RNA: forming intra strand pairing
  - Affects its stability, function (RNase), and others.

DNA Bending

- DNA structure is flexible.
- DNA double strand is bended by protein binding.
- The bending of the DNA structure will influence other proteins to bind to it.
Bend.it: Predict DNA Bendability

- Predicting DNA curvatures from DNA sequences
- The curvature is calculated as a vector sum of dinucleotide geometries (roll, tilt and twist angles).
- Expressed as degrees per helical turn.
- Service site: http://www2.icgeb.trieste.it/~dna/bend_it.html

RNA Draw: Predict RNA Secondary Structure on PC (1)

- Predicting RNA secondary conformation under given $E$ state
- Using dynamic programming, base pair probability and $E$ parameters

Clover leaf structure of tRNA and its 3D structure
**RNA Draw: Predict RNA Secondary Structure on PC (2)**

E.coli tRNA-Ile predicted structure on 37 C

---

**Protein Structure**

Primary structure (amino acid sequence) determine its secondary structure (partial structure: α helix, β-sheet) and tertiary structure

- Protein function: determined by its structure (e.g.: enzyme inactivated by heat - because its structure was changed by heat.)

(C) 2001 SNU CSE Artificial Intelligence Lab (SCAI)
Secondary Structure Prediction

- Prediction of secondary structure
  - $\alpha$-helix
  - $\beta$-sheet
- Can be helpful for its 3D conformation study.
- Solutions
  - Calculate “propensity” for each amino acid to be in helix, sheet…
  - Calculate “information content”
  - Neural network based method
  - Nearest neighbor

NNPREDICT: Protein Secondary Structure Prediction

- Two-layer, feed-forward neural network
- Finding secondary structure from amino acid sequences

http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
Tertiary Structure Prediction

- Using $E$ state?: only select lowest $E$ state - impossible
- Using homology modeling: similar sequence = similar structure

SWISS-MODEL

- Predicting protein 3D structure
Swiss-Pdb Viewer

SWISS-MODEL

5. DNA Microarray

DNA Microarray (1)

- Monitoring of gene expression
- cDNA chip: spotting cDNAs
- Oligonucleotide chip: spotting short DNA sequences
DNA Microarray: cDNA chip (2)

- DNAs are spotted on glass surfaces.
- Applications: drug effects, drug metabolism, disease diagnosis, finding gene pathway, finding disease genes, etc.

DNA Microarray: oligonucleotide chip (3)

- Short (10~25 nt) oligonucleotides on chip
- Applications: find sNPs, detect mutation, sequencing by chip, genotyping
DNA Microarray (4)

- Processing DNA chip images to obtain qualified data
- Data mining from DNA chip images like clustering (hierarchical algorithms, partitioning algorithms...)


- Useful websites:
  - www.affimatrix.com
  - www.genechip.co.kr
  - inkage.rockefeller.edu/wli/microarray
  - cmgm.stanford.edu/~plf58/microArray
6. Major Tools for Proteomics

Major Tools for Proteomics: 2D Gel & MALDI-TOF

- Proteomics:
  - Protein expression profile
  - Protein interaction
  - Protein structure
  - Protein variation

- Why proteomics?
  - Post-Genomics Era: Understanding functions
  - Applications in target discovery
  - Novel proteins and genes
  - Disease associated proteins
  - Elucidate pathways and processes in disease/cell biology
2-D Gel Electrophoresis

- Separates proteins by their netcharge and molecular weights.
- Displays whole proteins
- Like DNA chip, compares expression under different treatments, and detects tissues or cell specific expressions.
  - Finding specific proteins for further study
- Poor reproduction problem

2-D Gel Image of Human Liver
2-D Electrophoretic Gel Images

MALDI-TOF: Matrix-Assisted Laser Desorption Ionization

- Detection of protein molecular weights (mass spectrophotometer)
- Can determine protein sequence.
Other bioinformatic work

- Sequence assembly
  - From scattered DNA sequences, construct full sequence
- Specific database warehousing
  - Structural databases, protein block databases, drug databases
- sNPs
  - Find single nucleotide polymorphisms in population
- Construct gene to gene, protein to protein and gene to protein interaction pathway