Part I – Sequence analysis (DNA): Bioinformatics Software

Chen Xin
National University of Singapore

Bioinformatics software

Its role in research:

Bioinformatics software

- Cyclical refinement of predictive computer models used to define further biological experiments, including the optimization step.

- From Brusic et al. 2001, Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding, J. Mol. Graph. Model. 19:405-11, 467

Bioinformatics software

- By combining computational methods with experimental biology, major discoveries can be made faster and more efficiently.

- Today, every large molecular or systems biology project has a bioinformatics component.

- Use of biological software allows biologists to extend their set of skills for more efficient and more effective analysis of their data, and for planning of experiments.
Genetic information

- Genetic information carrier
  - DNA or RNA
- Genetic information carried
  - Sequence
- Hence:

\[
\text{Life} = f(\text{Sequence})
\]

New drug discovery

A drug =

- Target identification -> Lead discovery -> Lead optimization -> animal trial -> clinical trial

Target:
- Key to disease development
- Specific to disease development
- Sequence, Sample protein, 3D structure ...
DNA sequence analysis

Types of analysis:
- GC content
- Pattern analysis
- Translation (Open Reading Frame detection)
- Gene finding
- Mutation
- Primer design
- Restriction map
- ......

When you have a sequence

- Is it likely to be a gene?
- What is the possible expression level?
- What is the possible protein product?
- Can we get the protein product?
- Can we figure out the key residue in the protein product?
- ......
GC content

- Stability
  - GC: 3 hydrogen bonds
  - AT: 2 hydrogen bonds
- Codon preference
- GC rich fragment
  → Gene

GC Content

- CpG island
  - Resistance to methylation
  - Associated with genes which are frequently switched on
  - Estimate: ½ mammalian gene have CpG island
  - Most mammalian housekeeping genes have CpG island at 5’ end
GC content

- GC Content:
  - Emboss -> CompSeq
  - Emboss -> GEECEE
  - Bioedit

- CpG Island:
  - http://l25.itba.mi.cnr.it/genebin/wwwcpg.pl (Italy)
  - Emboss -> CpGReport

Pattern analysis

- Patterns in the sequence
- Associated with certain biological function
  - Transcription factor binding
  - Transcription starting
  - Transcription ending
  - Splicing
  - ......
Gene finding

- A kind of pattern search
- Gene structure
  - Promoter, Exon, Intron
  - Promoter: TATA box (TATAAT)
  - Exon: Open Reading Frame (ORF)
  - Intron: Only eukaryotes, have splicing signal
  - Other motifs

Gene

Picture from the LSM2104 Practical, V.B. LIT
Gene finding

- Most of the programs focused on Open reading frame
  - Emboss -> GetORF
  - Emboss -> ShowORF
- Other important elements:
  - Matrix binding site: Emboss -> MarScan
  - Promoter region: PromoterInspector
  - Splicing sites: GeneSplicer

Gene finding

- **Prokaryotes**
  - No intron
  - Long open reading frame
  - High density
  - Easy to detect
- **Eukaryotes**
  - Have intron
  - Combination of short open reading frames
  - Low density
  - Hard to detect
Problem 1:

- Is it a gene?
  - Not sure, but have some confidence
- What is the expression level if it is a gene?
  - Determined by the promoter and other upper stream elements

Translation

- Six reading frames
- Open reading frame (ORF)
  - Start codon
  - Stop codon
  - Certain length
- Tools: ShowORF
Conceptual translation

<table>
<thead>
<tr>
<th>Reading Frame</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>+2</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>+3</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>5'</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>3'</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
<tr>
<td>-1</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
<tr>
<td>-2</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
<tr>
<td>-3</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
</tbody>
</table>

Six reading frames

<table>
<thead>
<tr>
<th>Reading Frame</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>+2</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>+3</td>
<td>AATCCCAATCCCTCTGAGACATTCCCA</td>
</tr>
<tr>
<td>+2 (boxed)</td>
<td>M A I R V D * A</td>
</tr>
<tr>
<td>+3</td>
<td>W Q S A * T R</td>
</tr>
<tr>
<td>-1</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
<tr>
<td>-2</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
<tr>
<td>-3</td>
<td>TTAAGGTTAAGGCTACGAGATGT</td>
</tr>
</tbody>
</table>
Problem 2

- What is the possible product of this gene?
  - It is likely to be ....
  - This conceptual translation is in open reading frame ....

- Can we get the gene product?
  - If expression level high: Directly separate
  - If expression level low: Clone it

Transfer of the Insulin gene

Cloning the Insulin Gene

Transfer and cloning of the Insulin gene
Primer design

- Design primers only from accurate sequence data
- Restrict your search to regions that best reflect your goals
- Locate candidate primers
- Verification of your choice

Primer design

- (primer 1) CTAGTACGAT
- ATGCCGTAGATC......TCCGATCATGCTA
- TACGGCATCTAG......AGGCTAGTACGAT
- ATGCCGTAG (primer 2)
Primer design

- Mispriming areas
- Primer length: 18-30 (Usually)
- Annealing Temperature (55 - 75 C)
- GC content: 35% - 65% (usually)
- Avoid regions of secondary structure
- 100% complimentarity is not necessary
- Avoid self-complimentarity

Primer Design

**Online tools:**

- [http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-primer.html](http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-primer.html)
- [http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi](http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi)
- [http://www.cybergene.se/primer.html](http://www.cybergene.se/primer.html)

**Software tools**

- Omiga
- Vecter NTI
Restriction map

- Restriction enzyme
  - Recognize a pattern
  - Recognition site V.S. Cutting site
- Select restriction enzyme to get a fragment of sequence
- Rebuild the sequence to create or invalidate a restriction site
- Tools: Omiga, remap, bioedit
Mutation

- Can be generated by PCR
  - Primers that not perfectly match
- Frame shift mutation
  - Insertion
  - Deletion
- Substitution
  - Normal
  - Silent
Mutation

- Test the importance
  - Mutate suspected important place
- Create a pattern
  - Often silent mutation
- Invalidate a pattern
  - Often silent mutation
- Keep a reading frame

Problem 3

- Can we get the protein product?
  - Clone it and use a bacteria to express it
- Can we figure out the key residue in the protein product?
  - Guess the important residue
  - Mutate the residue to see whether the activity loses
Summary

- Life is determined by nucleotide sequences
- Sequence analysis reveals patterns have biological significance
- Sequence analysis helps the design of wet-lab experiments
- Next part will be on protein sequence analysis