Bioinformatics Methods

Iosif Vaisman
Department of Bioinformatics and Computational Biology

Email: ivaisman@gmu.edu

Spring 2012

Bioinformatics

Bioinformatics is a field that deals with biological information, data, and knowledge, and their storage, retrieval, management, and optimal use for problem solving and decision making.

NIH working definition of bioinformatics and computational biology (July 2000)

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Omics sciences

- Connectomics
- Cytomics
- Epigenomics
- Exposomics
- Exomics
- Genomics
- Glycomics
- Interferomics
- Interactomics
- Ionomics
- Kinomics
- Lipidomics
- Mechanomics
- Metabolomics
- Metagenomics
- Metallomics
- ORFeomics
- Organomics
- Pharmacogenomics
- Phenomics
- Physiomics
- Proteomics
- Regulomics
- Secretomics
- Speechomics
- Transcriptomics
Information and uncertainty

Information is a decrease in uncertainty

\[ \log_2(M) = -\log_2(M^*) = -\log_2(P) \]

Shannon's formula for uncertainty

\[ H = -\sum_{i=1}^{M} P_i \log_2 P_i \]

only infmrtn esntil to understand mst b tramnltd

Communicaton

Fundamental problem of communication:

reproducing at one point either exactly
or approximately a message selected at
another point

*The Mathematical Theory of Communication*
Claude Shannon and Warren Weaver
Communication system

INFORMATION SOURCE → ENCODER → SIGNAL → RECEIVED SIGNAL → MESSAGE → DESTINATION

Communication system duality

“This duality can be pursued further and is related to the duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.”

C. E. Shannon (1959)

Cell Informatics

DNA → mRNA → Protein

Sequence – structure – function

Luscombe et al., 2001
Error correcting codes

Hamming metric

The sum of bit changes necessary to move from one point in the permutation space to another point in the permutation space

0000 and 0111 are separated by Hamming distance of 3: 0000 - 0001 - 0011 - 0111

Information Theory

Nucleotide permutation space

Standard genetic code

1 bit

Griffiths et al., 2004
Differences from the Standard Code

Vertebrate Mitochondrial Code

<table>
<thead>
<tr>
<th>Codon</th>
<th>Yeast Mitochondrial Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGA</td>
<td>Ter * Arg R</td>
</tr>
<tr>
<td>AGG</td>
<td>Ter * Arg R</td>
</tr>
<tr>
<td>AUA</td>
<td>Met M Ile I</td>
</tr>
<tr>
<td>UGA</td>
<td>Trp W Ter *</td>
</tr>
</tbody>
</table>

Yeast Mitochondrial Code

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUA</td>
<td>Met M Ile I</td>
</tr>
<tr>
<td>CUU</td>
<td>Thr T Leu L</td>
</tr>
<tr>
<td>CUC</td>
<td>Thr T Leu L</td>
</tr>
<tr>
<td>CUA</td>
<td>Thr T Leu L</td>
</tr>
<tr>
<td>CUG</td>
<td>Thr T Leu L</td>
</tr>
<tr>
<td>UGA</td>
<td>Trp W Ter *</td>
</tr>
<tr>
<td>CGA</td>
<td>absent Arg R</td>
</tr>
<tr>
<td>CGC</td>
<td>absent Arg R</td>
</tr>
</tbody>
</table>

Noise Sources

- Vector sequences
- Heterologous sequences
- Rearranged & deleted sequences
- Repetitive element contamination
- Sequencing errors / Natural polymorphisms
- Frameshift errors

Frameshift Errors

5'3'Frame1 MKFGNPLLTVQPPPELSQTEVMKRLVN
5'3'Frame2 -NLETSSFLISHLSYLRP--SDWLI
5'3'Frame3 EIWKLPSHLSAT-AISDRSDEAIG-S
3'5'Frame1 RLTRFITSV--DSSGG--VRKFPNF
3'5'Frame2 D-PIASSLRSEIQAVDK-EGSFQIS
3'5'Frame3 INQSLHHFGLR--LAWLISEKEVSKFH

Comparative Sequence Sizes

- Yeast chromosome 3 350,000
- Escherichia coli (bacterium) genome 4,600,000
- Largest yeast chromosome now mapped 5,800,000
- Entire yeast genome 15,000,000
- Smallest human chromosome (Y) 50,000,000
- Largest human chromosome (1) 250,000,000
- Entire human genome 3,000,000,000

Standard genetic code

AAs = FVLLSSSSTYQLRSWKKKMGQIGQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQIQQ