Molecular Information Theory

R Rajesh

July 10, 2010
Ability to grow and reproduce.
Self replicate - both the genetic information (genome) and the structure carrying and protecting it (cell).
Growth and reproduction - information and energy in order to process raw materials into new living matter.
Create new organisms identical or very similar, to the original organism.
Evolution - descendents are not identical to their ancestors but gradually accumulate changes in their genetic information over time.
How Life is sustained?

- **GENETIC INFORMATION** - Biological Information is carried by nucleic acid molecules- DNA and RNA.
- Each units of genetic information is called GENE- physically nucleic acid molecules
- Total genetic information of an organism is called GENOME
- **ENERGY** - needed to put the genetic information to use- growth and reproduction.
- Processes in which energy is acquired, liberated and used for biosynthesis of cell components- METABOLISM.
Identity

- Synthesis of new cell components requires chemical machinery, e.g., ribosomes are needed for making proteins which make up the bulk of all living tissue.
- Characteristic outward physical form.
- Identity or self - not merely to assemble random organic material.
Main components of a cell

- **Membrane**: transport of nutrients, energy generation, cell signaling.
- **Genome**: storage of genetic information.
- **Cytoplasm**: site of most metabolic reaction.
- **Ribosome**: site of protein synthesis.
- Prokaryote - e.g. Bacteria; Eukaryote - Yeast, Mammals
- Cell contains many more components
Gene is portion of DNA
Chromosome contains DNA
DNA

- Double helix
- Inherent redundancy
- DNA- Sugar Phosphate and Base
- Name is from base - A, T, G, C
- Directional
- Base pairing rules : A-T and G-C
Is all of the DNA useful??

- Recent advances - some of the regions are useful - regulatory functions
Central Dogma of Molecular Biology

- DNA-mRNA-Protein
- mRNA read in pairs of three-codons

![DNA to Protein Sequence](image-url)
How it works??

- Transcription and Translation

![Diagram showing the process of transcription and translation in prokaryotes and eukaryotes.](image-url)
Protein

- Building blocks of proteins- Amino acids
- 20 Amino Acids
- eg Alanine (Ala), Glycine(Gly), Glutamic acid (Glu), Leucine (Leu).......and so on
- Protein is formed by chain of aminoacids- polypeptide chain
- 4 level structure- 3 dimesional structure is important to functionality
- Structural, Enzymes, Tranport, Regulatory.
The code book

- From codons (64)- Amino acid (20)- Protein

<table>
<thead>
<tr>
<th>2nd base in codon</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td></td>
<td>Phe</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phe</td>
<td>Ser</td>
<td>Tyr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leu</td>
<td>Ser</td>
<td>Cys</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td></td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td></td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Met</td>
<td></td>
<td>Arg</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leu</td>
<td>Pro</td>
<td>Arg</td>
</tr>
<tr>
<td></td>
<td>Leu</td>
<td>Met</td>
<td></td>
<td>Arg</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ile</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Met</td>
<td>Thr</td>
<td>Ser</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>Ala</td>
<td>Ala</td>
<td>Gly</td>
</tr>
</tbody>
</table>
Recap

- How life is sustained
Shannons communication system

- Theorems for source coding and channel coding
Biological communication system

Source alphabet

Source tape
Message in DNA alphabet
DNA tape
Nucleic RNA polymerase

Encoding
DNA alphabet to mRNA alphabet
Transcription

Channel code
Noise

Channel
mRNA message in RNA alphabet
Point mutation
Genetic noise

Channel
mRNA message + genetic noise

Decoding
Translation

Decoding
protein message in protein alphabet
Protein tape

Genetic noise
mischarged tRNA

Genetic noise
acylacylated tRNA

acyl synthetases
tRNA amino acids
Some thoughts

- Syntatic and Semantic 'information'
- Semantic information requires acquisition of Syntatic data.
- Shannon’s theory: managing information only from a purely syntactic perspective
- Concept of information plays a central role in all the profound aspects pertaining to molecular biology
- Methods, ideas and tools of Information theory - gain insights about statistical regularities of data derived from biological sequences of nucleotides or amino acids.
- Information has a nature that does not depend on the material substrates over which it is generated and transmitted, being silicon systems or biochemical systems
- Information laws are universal
- Discover the secrets of Molecular biology, its wonderful working, in terms of information transmission, conservation and correction.
Medical application - concerned with the fine structure of DNA.

Structure of the message sent on the biomolecular channel

Little change in a nucleotide can dramatically change the function of a protein - destruction of the corresponding biological organism.

Employment of IT methods are confined to those cases in which a statistical analysis is part of the framework - biological insights.
Amino acids - 20
Number of $N$ length sequences - $20^N$
Assumption is that all the amino acids are equi-probable
- Often not the case
Recently information theoretic elements like - Asymptotic Equi partition property is applied
Number of sequences are found using $2^{NH}$. 
Genetic Channel

- Sender is specified by $[\Omega, A, p_A]$ and Receiver by $[\Omega, B, p_B]$
- Transition probabilities- All the effects lumped in the channel
- Biological cause- Mutations
- If interchange of letters is made with equal probability - white genetic noise
- Burst noise- e.g an $\alpha$ particle passing through DNA
DNA to Protein Bio-molecular channel

- Error correction

- Doesn't look efficient
- Error detection and correction capabilities
- Best protected is *Leu* and *Arg* and least protected is *Trp*.
- The stability of biological organisms shows that in practice synergism among various techniques used by Nature is adequate to the scope
DNA to Protein Bio-molecular channel

- Model the effect of the noise distributed over the DNA to Protein communication channel.
- Model noise as mischarged tRNA noise - decoding error
- Transition probability matrix - Assumption: Exchange probability is same
- CUG codes for Leu
- CUG can become Gln, Pro, Arg, Met and Val with probability $p$ miscoding correctly decoded as Leu with $1 - p$
- More realistic model required - exact probabilities can be calculated and modelling the various sources of noise.
- Capacity of the bio-molecular channel (biological sense) and the evaluation of the error correcting capabilities of the bio-code.
Mutual Information in Biology

- Measure of similarity of degree of homology was given as a percentage score.
- Recently mutual information based characterizations available.
- $I(A; B)$ is the new measure.
- Can also measure information content/complexity of a protein family.
Finding splicing sites

- Consider a Eularyotic gene
Finding splicing sites

- Points where introns are called splicing sites
- Identified by Patterns
- Each pattern is characterized by a specific conserved sequence motif (consensus sequence)
- These conserved motifs are expected to be quite complex and very stable under evolutionary pressure
- Basis for searching homology among (sequences derived from) different genes and/or different organisms
- Sequence Alignment
Finding splicing sites

- Finding consensus sequence.
- Solved using conditional entropy and a threshold based rule.
Site redundancy for locating Binding sites

- Using mutual information $H_{\text{before}} - H_{\text{after}}$
- Prior probabilities -[A,T,G,C] equiprobable
- Take the example of initiation codon (AUG/GUG) rarely it can be UUG or CUG
- Information gained by the ribosome for the second position is 2 bits
Site redundancy for locating Binding sites

\[ H_{\text{after}}(l) = - \sum_{b \in \{A,T,G,C\}} f(b, l) \log f(b, l) \text{ bits/base} \]

\[ H_{\text{before}} = 2 \text{bits/base} \]

\[ R_{\text{sequence}}(l) = H_{\text{before}} - H_{\text{after}} \]

\[ R_{\text{sequence}} = \sum_{l} R_{\text{sequence}}(l) \text{ bits/site} \]

\[ R_{\text{sequence}} \text{ gives the total information.} \]
Site redundancy for locating Binding sites

- Find the SD (Shine-Dalgarno)
- Also called ribosome binding site (RBS)
Site redundancy for locating Binding sites
Discriminating between coding and non-coding regions

- Several techniques were developed to extract information about the (coding/non-coding) status of the DNA
- Using mutual information
Conclusion

- Used in many applications
- mainly by biologists
- More accurate modelling can answer more questions