ECEA 2019
5th International Electronic Conference on Entropy and Its Applications
18–30 November 2019
Chaired by Prof. Geert Verdoolaege
Sponsored by: entropy MDPI

Information Theory in Computational Biology

Pritam Chanda, PhD
Research Scientist, Data Science & Informatics,
Corteva AgriScience R&D, Indianapolis, IN, USA

Talk Organization

- Information theoretic measures : Definitions and terminology
- Gene regulatory network inference
- Identifying disease associated genetic variations
- Biological Sequence Analysis: alignment free phylogeny

Information Theory

- "A Mathematical Theory of Communication" - Claude Shannon (1948).
  - Data transmission through (noisy) channels

- Diverse applications : physics, computer science, statistics, economics, neurobiology, genetics, epidemiology, ecology, bioinformatics and computational biology.

- Information theory and the living system – Lila L Gatlin, 1972
  - Information content of DNA. (J. Theor. Biol. 1966)
  - Information content of DNA. II. (J. Theor. Biol. 1968)

Information theoretic measures

- Entropy $$H(f) = -E[\log f(x)]$$

Let X be a random variable which takes its values from χ and its probability mass function be:

$$p(x) = Pr\{X = x\}, \quad x \in \chi$$

$$H(X) = -\sum_{x} p(x) \log p(x) = E[-\log p(x)]$$

- Joint Entropy

$$H(X, Y) = -\sum_{x \in \chi} \sum_{y \in Y} p(x, y). \log p(x, y)$$

- Relative Entropy (KLD)

$$D(p||q) = \sum_{x \in \chi} p(x). \log \frac{p(x)}{q(x)}$$

Information theoretic measures

- Mutual Information (MI)

$$I(X;Y) = \sum_{x \in \chi} \sum_{y \in Y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)}$$
$$= D(p(x,y) || p(x).p(y))$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- Conditional Mutual Information (CMI)

$$I(X,Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

$I(X;Y|Z) - I(X;Y)$, is called Interaction information

Information theoretic measures

- K-Way interaction Information (KWII)
  - Amount of information (synergy or redundancy) that is present in the set of variables, which is not present in any subset of these variables

$$KWII(X_1;X_2;X_3) = -H(X_1) - H(X_2) - H(X_3) + H(X_1, X_2) + H(X_1, X_3) + H(X_2, X_3) - H(X_1, X_2, X_3)$$

For the n-variable case on the set $v = \{X_1, X_2, \ldots, X_n\}$

$$\mathrm{KWII}(v) \equiv -\sum_{T \subseteq v} (-1)^{|v|-|T|} H(T)$$

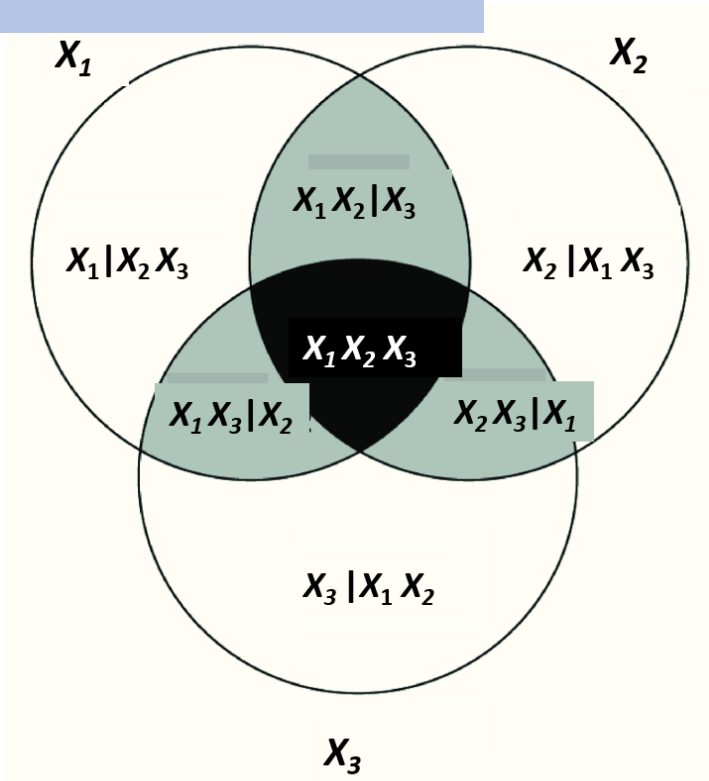
Information theoretic measures

Total Correlation Information: Total amount of information shared among the variables in the set.

$$\mathrm{TCI}(X_1, X_2, \ldots, X_n) = \left[\sum_{i=1}^{n} H(X_i)\right] - H(X_1, X_2, \ldots, X_n)$$

$$TCI(X_1, X_2, \ldots, X_n) = \sum_{\nu \in \{X_1, X_2, \ldots, X_n\}, |\nu| \geq 2} KWII(\nu)$$

Phenotype Association Information: Total amount of information shared among the variables in the set with respect to a class variable.



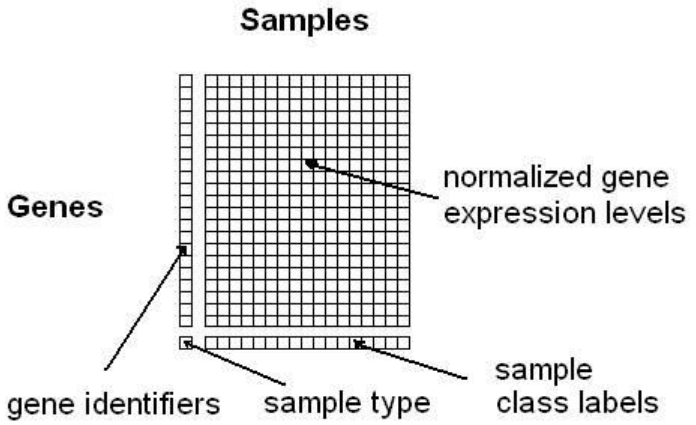
Bell AJ at. al. Proc. 4th Int. Symp. Independent Component Analysis and Blind Source Separation, 2003

Chanda et al. Am J Hum Genet. 2007, 81 (5): 939-963

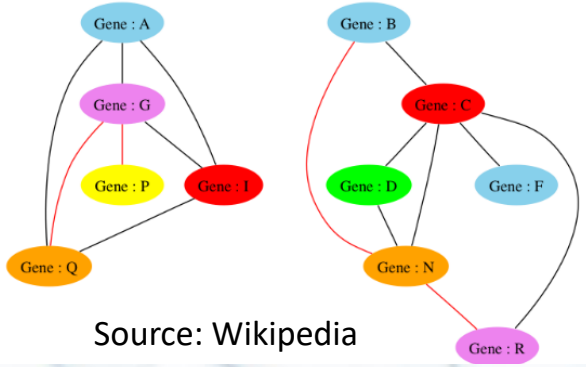
$$PAI(X_1, X_2, \ldots, X_K, P) = TCI(X_1, X_2, \ldots, X_K, P) - TCI(X_1, X_2, \ldots, X_K)$$

Gene Regulatory Network Inference

- Reverse engineering transcriptional regulatory network from expression data
  - Genome-wide clustering of gene expression profiles
    - Coarse representation of genes that are co-regulated together
  - Gene-gene interaction network/graph from expression data
    - A graph G(V, E) represents a network where V denotes a set of genes and E denotes a set of regulatory relationships between genes.
    - Each gene/transcription factor is a node in the network/graph
    - Each edge models the statistical dependency between the two nodes.
    - If gene x shares a regulator relationship with gene y, then there exists an edge between x and y, i.e (x --- y).



"Is there a regulatory interaction from gene X to gene Y



Source: Wikipedia

Mutual Information (MI) between genes

$$I(G_i, G_j) = \sum_{g_i \in \Omega} \sum_{g_i \in \Omega} p(g_i, g_j) log \left(\frac{p(g_i, g_j)}{p(g_i)p(g_j)} \right)$$

between the gene random variables G_i and G_j

Pairwise measurements for every pair of genes in the expression matrix

Mutual information is zero if G_i and G_j are independent

Relevance Networks : if the mutual information between the expression levels of two genes is higher than a threshold, it is more likely that they have a biological relationship

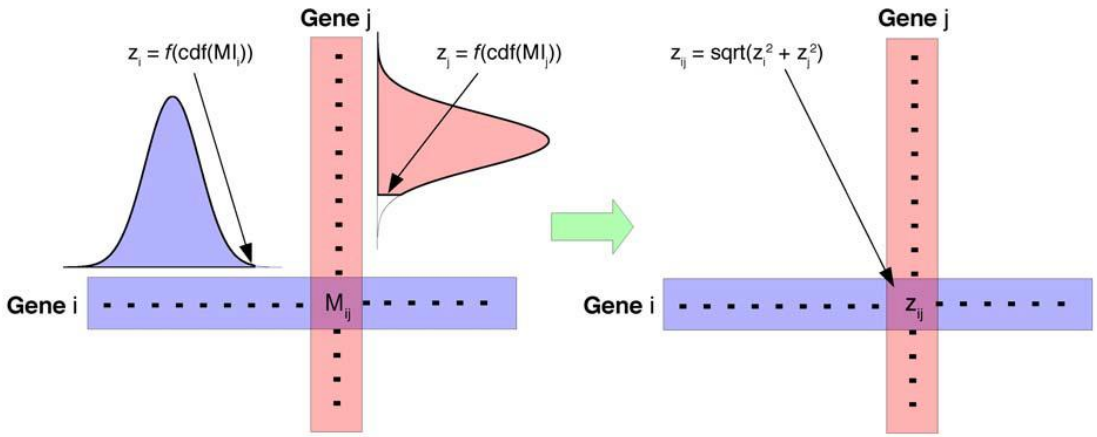
edge iff : $I(G_i;G_j) \geq$ threshold

A. J. Butte and I. S. Kohane, Pacific Symposium on Biocomputing, pp. 418–429, 2000.

CLR (context likelihood of relatedness)

Faith JJ, et al. PLoS Biol. 2007;5(1):e8.

Considers the background distribution of the MI values



$$z_i = \max_j \left(0, \frac{I(G_i;G_j) - \mu_i}{\sigma_i}\right)$$

Mean and standard deviation of MI values { $I(G_i; G_k)$ }

$k = 1, \ldots, N$

$$score_{i,j} = \sqrt{z_i^2 + z_j^2}$$

Compares the MI between a pair of genes G_i and G_j to the background distribution of mutual information scores for all possible gene pairs that include either G_i or G_j .

At a 60% true positive rate, CLR identified 1,079 regulatory interactions (741 novel predictions) in E coli.

ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks)

A. A. Margolin, et al., BMC Bioinformatics, vol. 7, supplement 1, p. S7, 2006.

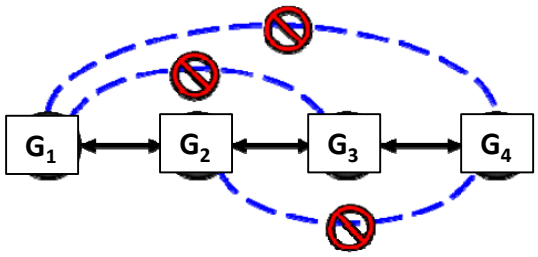
Compute pairwise Mutual Information for every pair of genes

- Estimated as $\hat{I}(G_i;G_j) = \frac{1}{M}\sum_i log\left(\frac{f(G_i,G_j)}{f(G_i)f(G_j)}\right)$
- f(.) estimated using Gaussian Kernel Density estimation

Filter Interactions

- $I(G_i;G_j) > I_{th}$ are only retained.
- I_{th} : shuffling the gene expressions to get null distribution of mutual information

Data Processing Inequality : $I(G_i;G_j) \leq min(I(G_i;G_k), I(G_k;G_j))$



G_1, G_2, G_3, and G_4 are connected in a linear chain relationship.
$I(G_1;G_2) > I(G_1;G_3)$ and $I(G_2,G_3) > I(G_1,G_3)$: remove G_1--- G_3
$I(G_2, G_3) > I(G_2, G_4)$ and $I(G_3, G_4) > I(G_2, G_4)$: remove G_2 --- G_4

$I(G_1, G_2) > I(G_1, G_4)$ and $I(G_2, G_4) > I(G_1, G_4)$, : remove G_1--- G_4
$I(G_1, G_3) > I(G_1, G_4)$ and $I(G_3, G_4) > I(G_1, G_4)$.

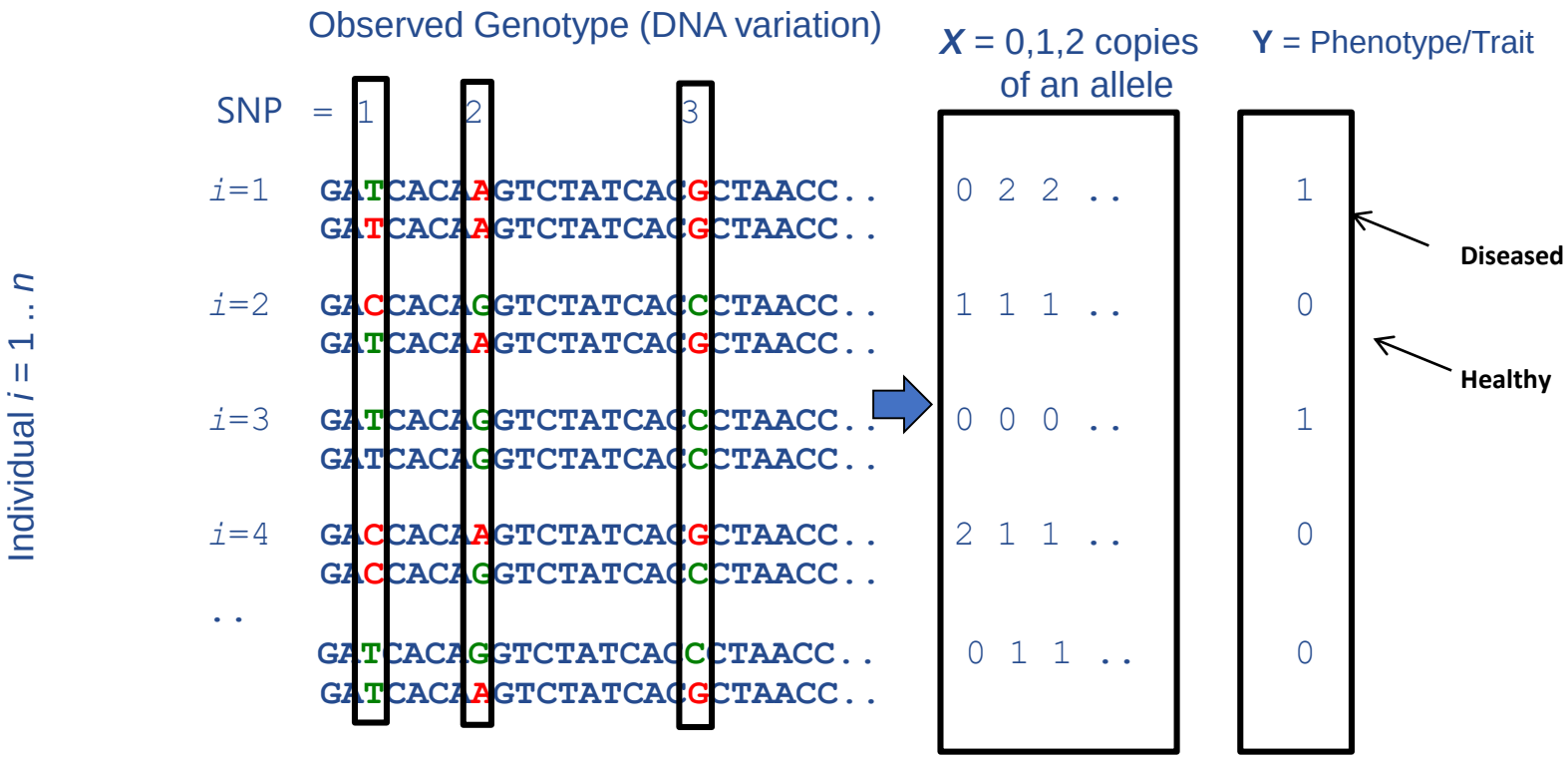
Other notable methods

- MRNET : based on maximum relevance/minimum redundancy (MRMR) principle [Meyer PE, et al. EURASIP J. Bioinf. Syst. Biol 2007.]

- Predictive Minimum Description Length (PMDL)
  - Used minimum description length (MDL) to find a threshold for MI, then CMI to infer regulatory relationships [Chaitankar V, et al. BMC Syst. Biol 2010;4(Suppl 1):S7.]

- PCA-CMI : higher order CMI and path consistency algorithm to prune a MI based gene network [Zhang X, et al. Bioinformatics 2012;28(1):98–104.]

Identifying Disease Associated Statistical Interactions

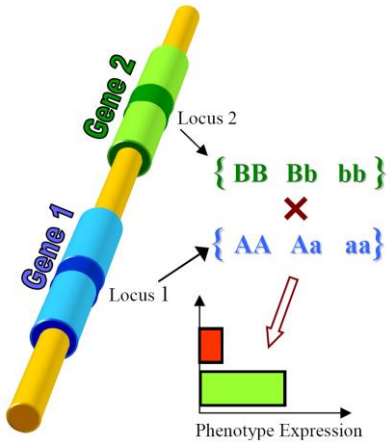


Total genome length = 3×10^9

Common biallelic variable sites = SNP ~ 10 million

Gene-Gene Interactions (GGI) and Gene-Environment Interactions(GEI)

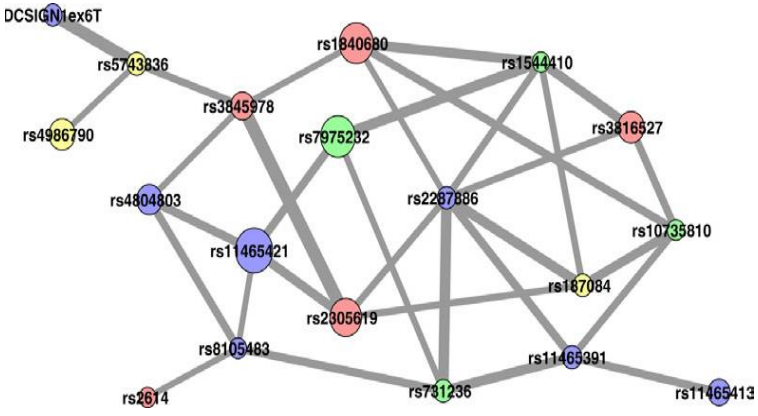
- Analyzed and visualized using KWII and TCI. [Chanda et al. Am J Hum Genet. 2007, 81 (5): 939-963]
- AMBIENCE: greedy algorithm to identify GEI and GGI in genome-wide data using KWII and PAI [Chanda et. Al (2008) Genetics. 2008, 180 (2)].
- Higher power in identifying interactions [Sucheston et. al. BMC Genomics , 2010, vol. 3 pg. 487]
- CHORUS: Higher order interaction identification algorithm for quatitative traits [P Chanda, BMC genomics 10 (1), 509]



Epistasis networks

- Graph G={V,E}. V = {SNPs}. E = {Edges between SNPs}.
- Initial weight for each SNP using MI : $I(SNP_i,P)$, P = Bladder cancer susceptibility. [Hu T, 2011 BMC Bioinformatics 12:364].
- Edge between SNP_i and SNP_j : $KWII(SNP_i;SNP_j;P) \geq$ threshold
- Permutations to assess threshold and significance.

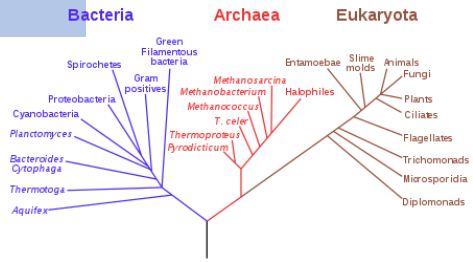
- Network Properties: connected components size, vertex degree distributions.
  - Approximately scale free.
  - Multi-SNP disease associations (connected components of size>=2)
  - existence of main effects does not necessarily correlate with the occurrence of interactions.



Hu et al. J Am Med Inform Assoc 20:630–636

Alignment free genome comparison

- Why alignment free ?
  - Assumption that similar sequences will have conserved sequence stretches is often violated.
  - Accuracy of sequence alignment methods drops off rapidly when the sequence identity falls below a certain critical point.
  - Computationally intensive and memory constrained for use with multi-genome-scale sequence data.
  - Alignment methods often have arbitrary parameters (gap penalty, various substitution matrices etc.).
  - Often difficult to align genomes, each species can have its own gene content.



A phylogenetic tree

Source: Wikipedia

- Alignment-free approaches to sequence comparison :

  method of quantifying sequence similarity that does not use or produce alignment (assignment of residue–residue correspondence) at any step of algorithm application

Alignment free genome comparison using feature frequency profiles

Sims GE et al. Proc Natl Acad Sci U S A. 2009 Feb 24;106(8):2677-82.

- Determining the similarity/dissimilarity between a pair of genomes
- "Words" from genome – sliding window of length l from position 1 to $n - l + 1$

A G G G T A A A A C T G T G T G C C A A A T G C

Sliding windows of length 9
words = { AGGGTAAAA, GGGTAAAAC, … , GCCAAATGC}

- Count the frequency of each "word" (l-mer)
- Feature Frequency Profile (FFP) for a genomic sequence
- Distance between two genomic FFPs P_l and Q_l :

Count of a l-mer

4^l

$$F_l = \frac{< c_{l,1}, c_{l,2}, \cdots c_{l,K} >}{\sum_i c_{l,j}}$$

Jensen Shannon Divergence (JSD)

$$JSD(P_l, Q_l) = \frac{KL(P_l, M_l) + KL(Q_l, M_l)}{2} \qquad M_l = \frac{P_l + Q_l}{2}$$

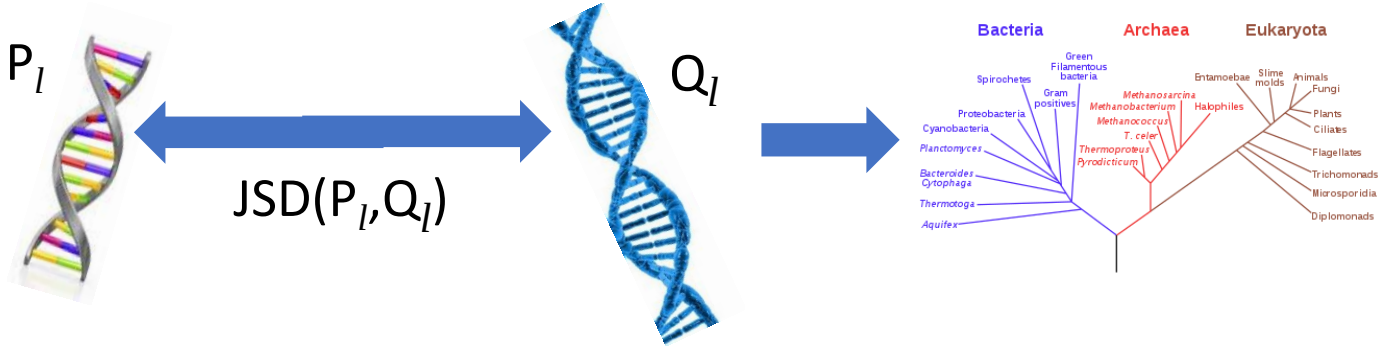
Symmetric

$$KL(P_l, M_l) = \sum_{i=1}^{K} p_{ij} log \frac{p_{ij}}{m_{li}}$$

Kullback Leibler Divergence (KLD)

Asymmetric

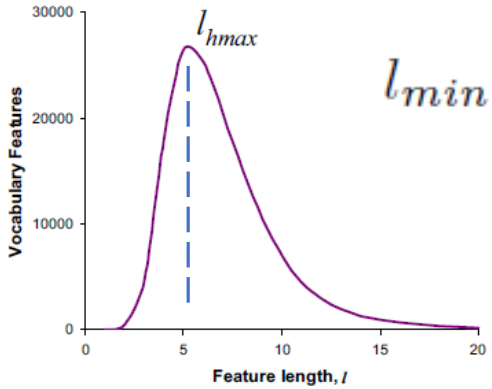
Alignment free genome comparison using feature frequency profiles



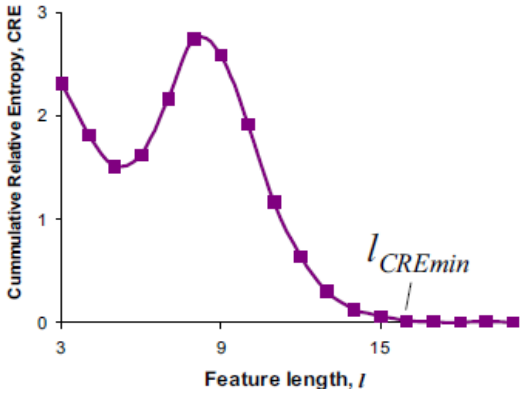
How to find optimal *l* ?

$$CRE(l) = \sum_{i=1}^{\infty} |KL(\hat{F}_i, F_i)|$$

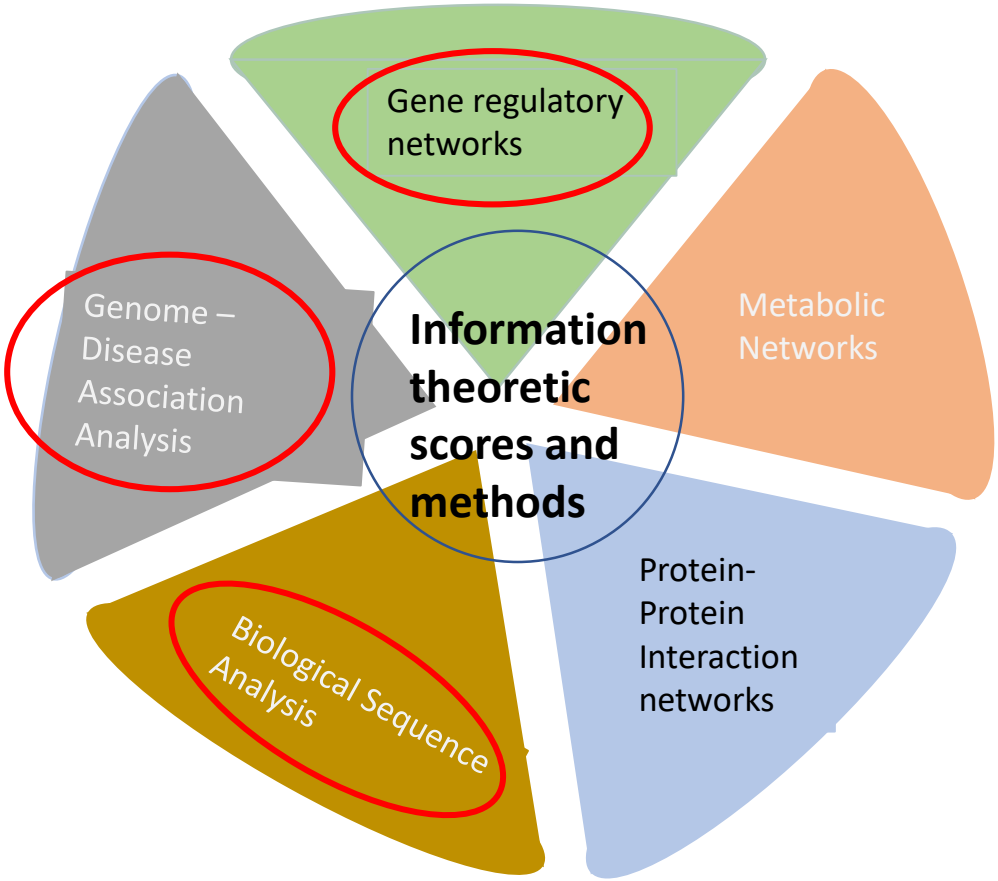
$$\hat{f}_{a_1 a_2 \cdots a_l} = f_{a_1 a_2 \cdots a_{l-1}} \frac{f_{a_2 \cdots a_l}}{f_{a_2 \cdots a_{l-1}}}$$



$$l_{min} \leq l \leq l_{max}$$



Many applications in Bioinformatics, Computational and Systems Biology



- Entropy
- Joint Entropy
- Mutual Information
- Relative Entropy (KLD)
- Conditional Mutual Information
- Total Correlation Information
- K-Way Interaction Information
- Jensen-Shannon Divergence
- Phenotype Associated Information

Thank You!