Mutual information and its applications

Lab meeting – 11/2/2011

Vikas Rao Pejaver

Preliminaries

- For a random variable X ~ p(x), Shannon's entropy:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

- Entropy is a measure of uncertainty of a random variable: when p(x) = 1, H(X) = 0

- Also, average information content missing when the value of the random variable is not known

- Relative entropy: $D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$

$$\begin{array}{rcccc} D_{\mathrm{KL}}(P\|Q) & = & -\sum_x p(x) \log q(x) & + & \sum_x p(x) \log p(x) \\ & = & H(P,Q) & - & H(P) \end{array}$$

Cross-entropy **Entropy**

What is mutual information?

- For random variables X and Y, it is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

- KL divergence interpretation: The penalty when X and Y are assumed to be independent
- Mutual information (MI) measures the dependence of X and Y
- Information gain on X once Y is known:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

MI – a measure of covariance

- If X and Y are independent, $p(x,y) = p_1(x).p_2(y)$

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right) = 0$$

- Higher the value of p(x,y), higher the MI
- Pearson product-moment correlation co-efficient:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

- Most correlation measures including Pearson's correlation quantify only a linear dependence between variables
- Information theory provides a more general measure of dependencies

MI – a distance metric?

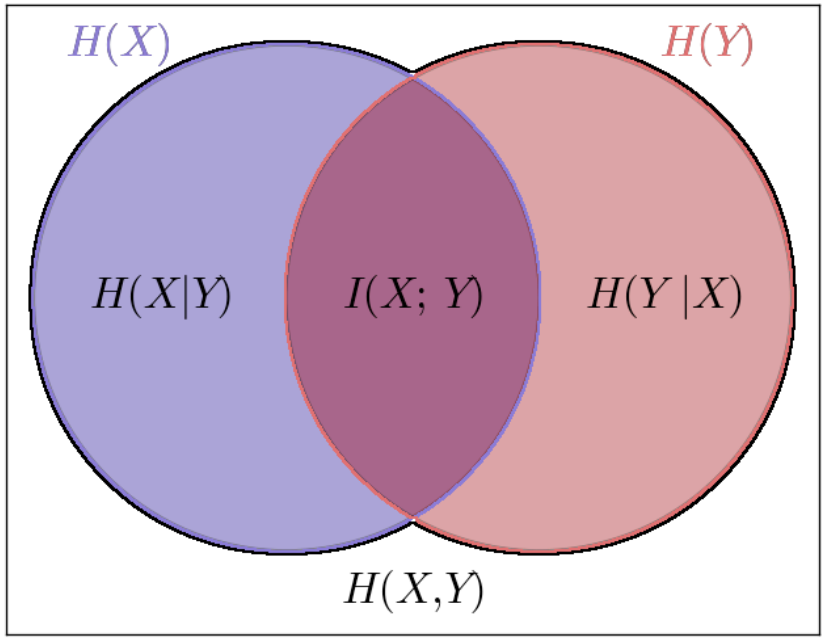
- MI by itself is not a metric (because KL divergence is not a metric)
- So use 'variation of information':

$$\mathbf{d(X,Y) = H(X,Y) - I(X;Y)} = H(X) + H(Y) - 2I(X;Y)$$
$$= H(X \mid Y) + H(Y \mid X)$$

- Satisfies the properties of a metric: triangle inequality, non-negativity, indiscernability and symmetry
- A natural normalized variant: $D(X,Y) = d(X,Y)/H(X,Y) \leq 1.$
- D is a universal metric: if any other distance measure places X and Y close by, then so will D

MI – a distance metric?

- Now, D(X,Y) = 1 – I(X;Y) / H(X,Y)



which is the Jaccard distance!

Variants of mutual information

- Co-efficients of constraint or uncertainty co-effs:

$$C_{XY} = \frac{I(X;Y)}{H(Y)} \quad \text{and} \quad C_{YX} = \frac{I(X;Y)}{H(X)}.$$

- Redundancy:

$$R = \frac{I(X;Y)}{H(X) + H(Y)}$$

- Symmetric uncertainty:

$$U(X,Y) = 2R = 2\frac{I(X;Y)}{H(X) + H(Y)}$$

- Weighted variants:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} w(x,y) p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)},$$

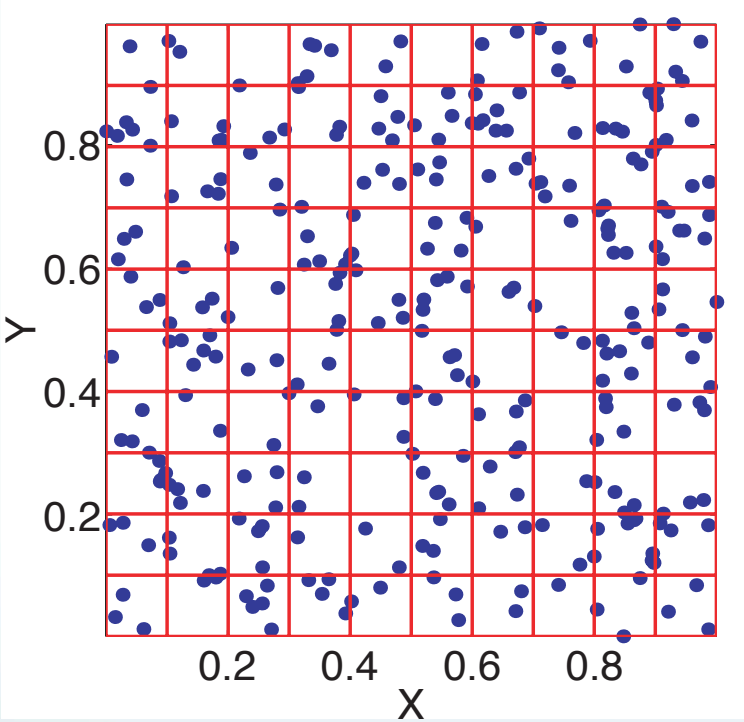
But how does one estimate MI?

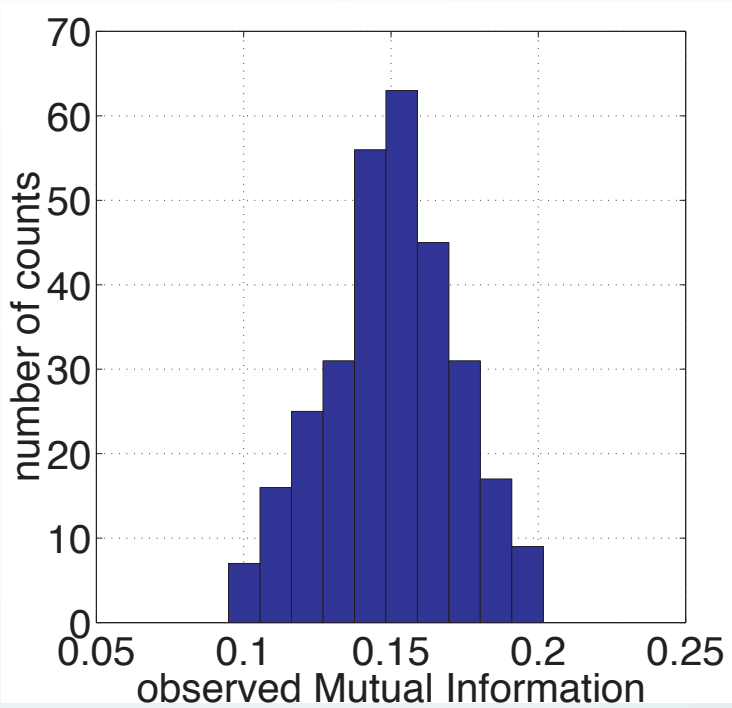
- In the discrete case, it is pretty straightforward
- However, in the continuous case, assigning probabilities to ‘events’ is not intuitive
- Numerical estimation:
  - Bin your values and calculate probabilities for each bin
  - Proceed as discussed before
- Kernel density estimation:
  - Use some sort of an estimator to get densities
  - Integrate over the smooth density function

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

Numerical estimation

- Naïve algorithm:





$$I(X,Y) = \log N + \frac{1}{N}\sum_{ij} k_{ij} \log \frac{k_{ij}}{k_i k_j}$$

But since X and Y are independent, MI must be 0!

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

Numerical estimation

- Finite size effects (Herzel et al.): $\langle H^{observed} \rangle \approx H - \frac{M-1}{2N}$

$$\langle I^{observed} \rangle \approx I(X,Y)^{\text{true}} + \Delta I(X,Y)$$

$$\Delta I(X,Y) = \frac{M_{xy} - M_x - M_y + 1}{2N}$$

- Adaptive partitioning:
  - Instead of using fixed intervals, divide the axes such that the width of each interval is determined by the local density of the measured dataset
  - Faser-Swinney algorithm constructs a hierarchy of partitions
  - But not much better than basic adaptive partitioning

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

Kernel density estimation

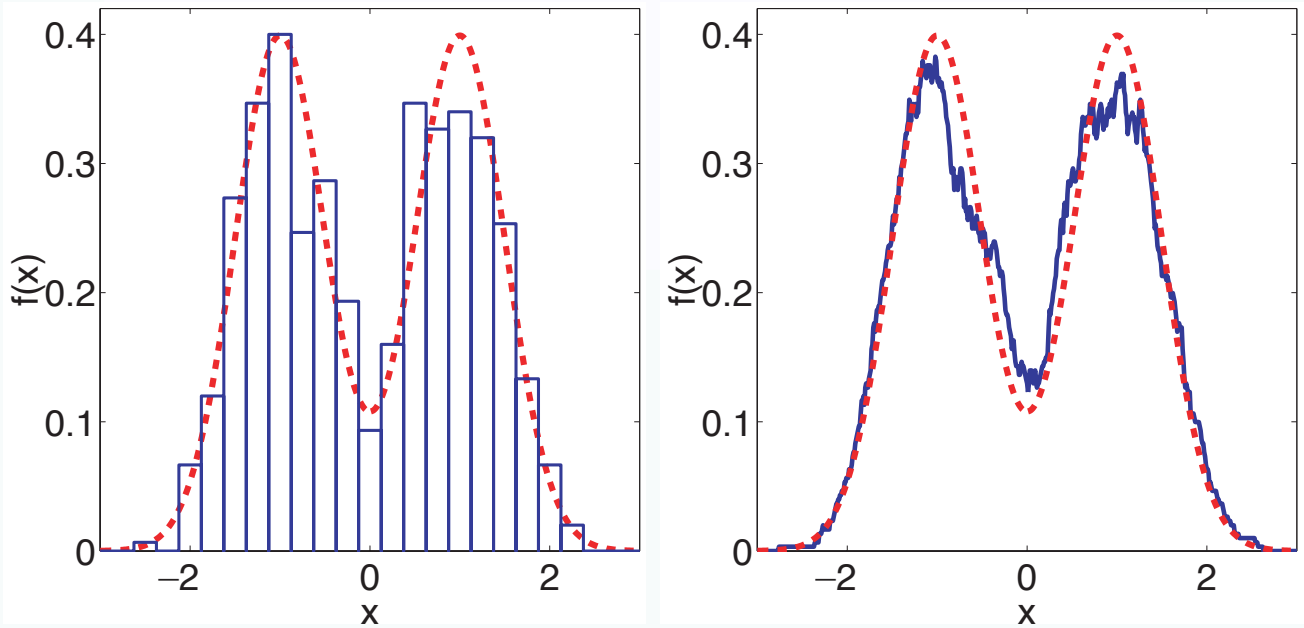
- Superior to histograms method:
  - Better mean square error rate of convergence of the estimate to the underlying density
  - Insensitivity to choice of origin
  - Ability to specify more sophisticated window shapes than a rectangular window
- First, free the histogram from a choice of origin and bin position - naïve estimator

$$\hat{f}(x) = \frac{1}{2Nh} \sum_{i=1}^{N} \Theta(h - |x - x_i|)$$

$$\Theta(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}$$

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

Kernel density estimation



- With a generalized weight or kernel function K(x), we get the kernel density estimator:

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

where h is a smoothing parameter or window width

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

Kernel density estimation

- What about 2D case? Think of this example – a Gaussian kernel in 1D:

$$\hat{f}(x) = \frac{1}{N}\frac{1}{h\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2h^2}\right)$$

- For 2D, you just take the Euclidean distance d(x,y):

$$\hat{f}_g(x,y) = \frac{1}{Nh^2}\frac{1}{2\pi}\sum_{i=1}^{N}\exp\left(-\frac{d_i(x,y)^2}{2h^2}\right)$$

- Choice of h is crucial:
  - If h is too small, spurious fine structure is visible
  - If h is too large, details will be missed out
- There are methods to estimate optimal h

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

Kernel density estimation

- In our example:

$$h_{opt} \approx \sigma \left(\frac{4}{d+2} \right)^{1/(d+4)} N^{-1/(d+4)}$$

  where d is the dimension of the dataset

- Now, to get MI, you just integrate:

$$\hat{I}(X,Y) = \int_x \int_y \hat{f}(x,y) \log \frac{\hat{f}(x,y)}{\hat{f}(x)\hat{f}(y)} \mathrm{d}x \, \mathrm{d}y$$

- A simpler approximation:

$$\hat{I}(X,Y) = \left\langle \log \frac{\hat{f}(x,y)}{\hat{f}(x)\hat{f}(y)} \right\rangle \Rightarrow \hat{I}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{\hat{f}(x_i,y_i)}{\hat{f}(x_i)\hat{f}(y_i)} \right]$$

Steuer et al. 2002. Bioinformatics. 18(2), 231-240.

MI – Identifying correlated residues

$$C_{ij} = C_{\text{phylogeny}} + C_{\text{structure}} + C_{\text{function}} + C_{\text{interactions}} + C_{\text{stochastic}}.$$

- Assume X (a site) is a discrete random variable for which there is uncertainty about the 20 values it could take (amino acids)
- But expected frequencies are known for each amino acid
- We can calculate how much information is present at each site:

$$E(X) := -\sum_{j=1}^{n} p_j \log_2(p_j),$$

- Min. = 0 and Max. = $\log_2(n)$ [Uniform distribution]

Atchley et al. 2000. Molecular Biology and Evolution. 17(1), 164-178.

MI – Identifying correlated residues

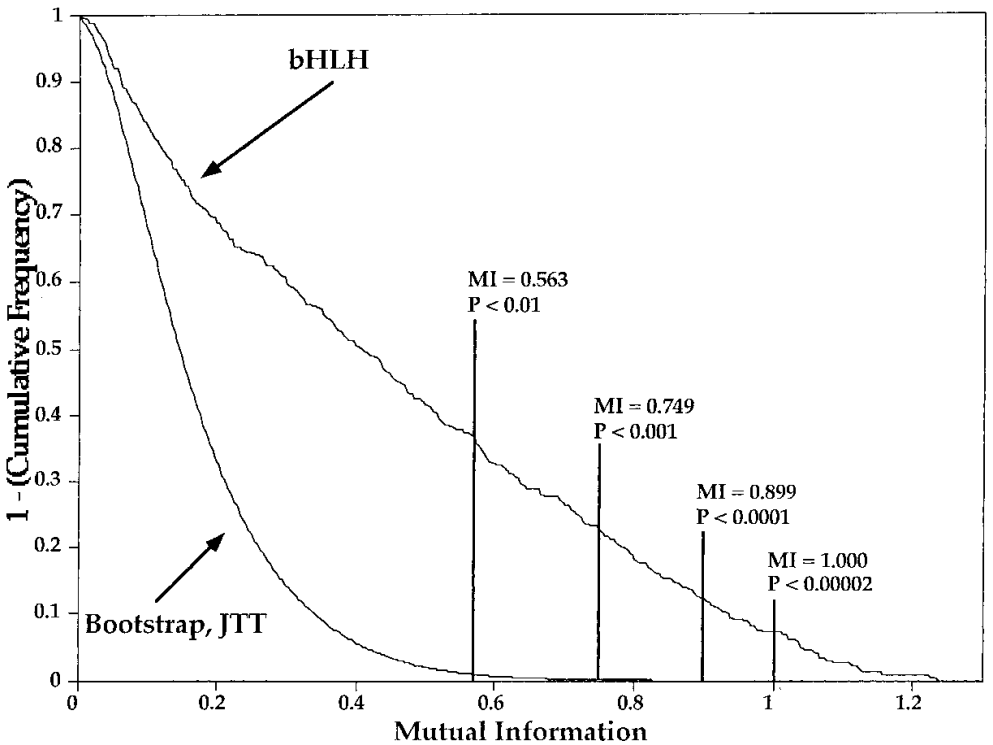
- MI for a pair of sites X and Y provides the relative information content of Y contained in X:

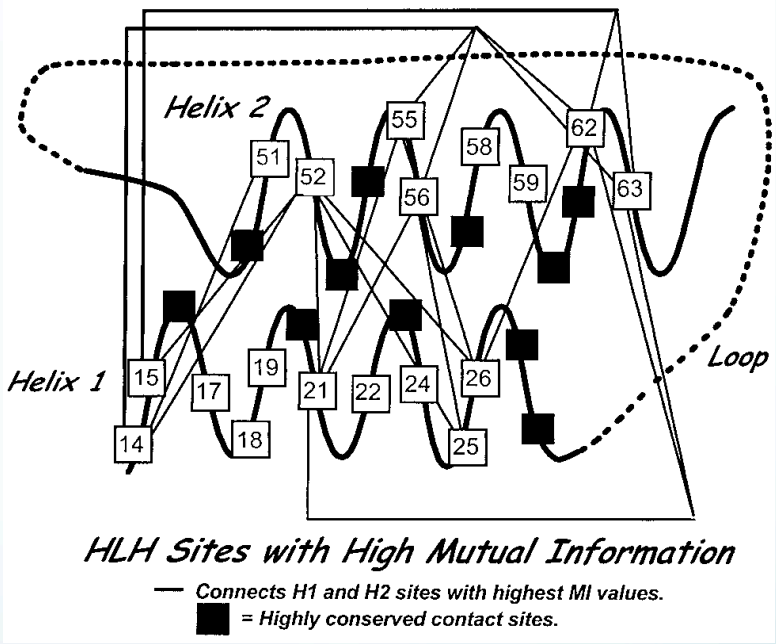
$$MI(X, Y) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_{jk} \log_2 \frac{p_{jk}}{p_j q_k}$$

- Used a parametric bootstrap and generated a distribution of MI values for these bootstrap datasets
- Statistical significance of MI values determined by comparing frequency distributions of MI values over all 237 bHLH sequences and the bootstraps
- Isolated correlated sites purely due to structure, function and interaction

Atchley et al. 2000. Molecular Biology and Evolution. 17(1), 164-178.

MI – Identifying correlated residues





Atchley et al. 2000. Molecular Biology and Evolution. 17(1), 164-178.

MI – Clustering gene expression data

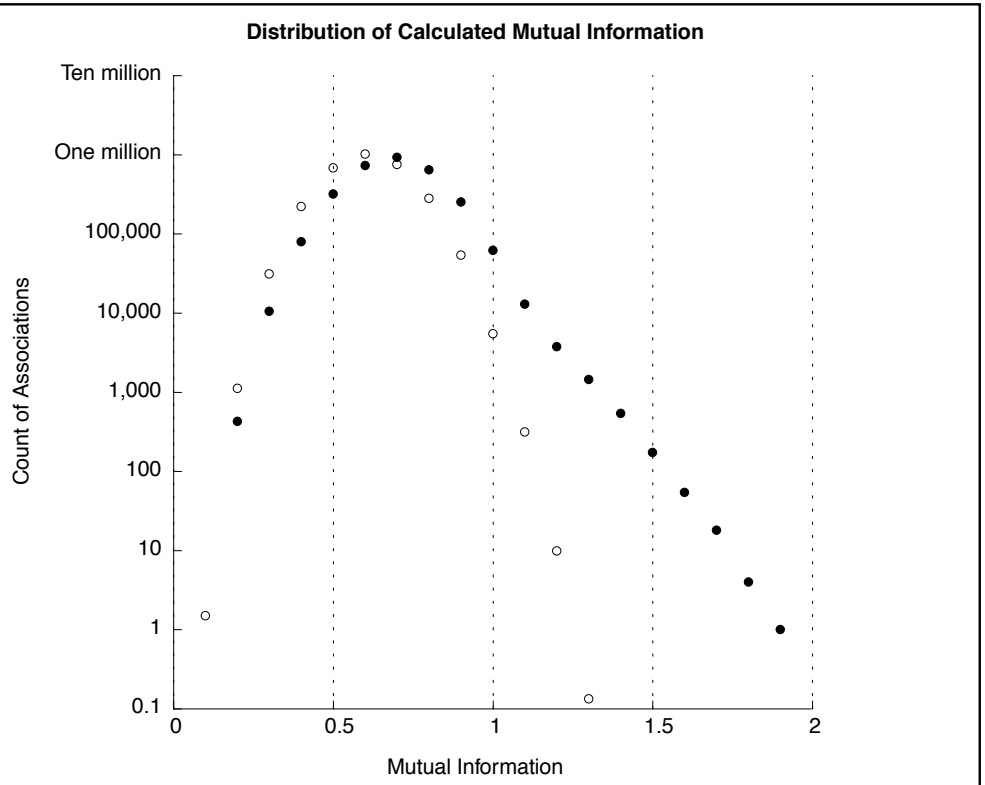


Figure 1: Pair-wise mutual information was calculated between 79 measurements of RNA expression of 2,467 genes in *Saccharomyces cerevisiae* and the distribution of these is shown with filled circles. The same was calculated using permuted RNA expression measurements; the average distribution from 30 permuted repetitions is shown with open circles.

Butte and Kohane. 2000. Pacific Symposium in Biocomputing. 418-429.

MI – Clustering gene expression data

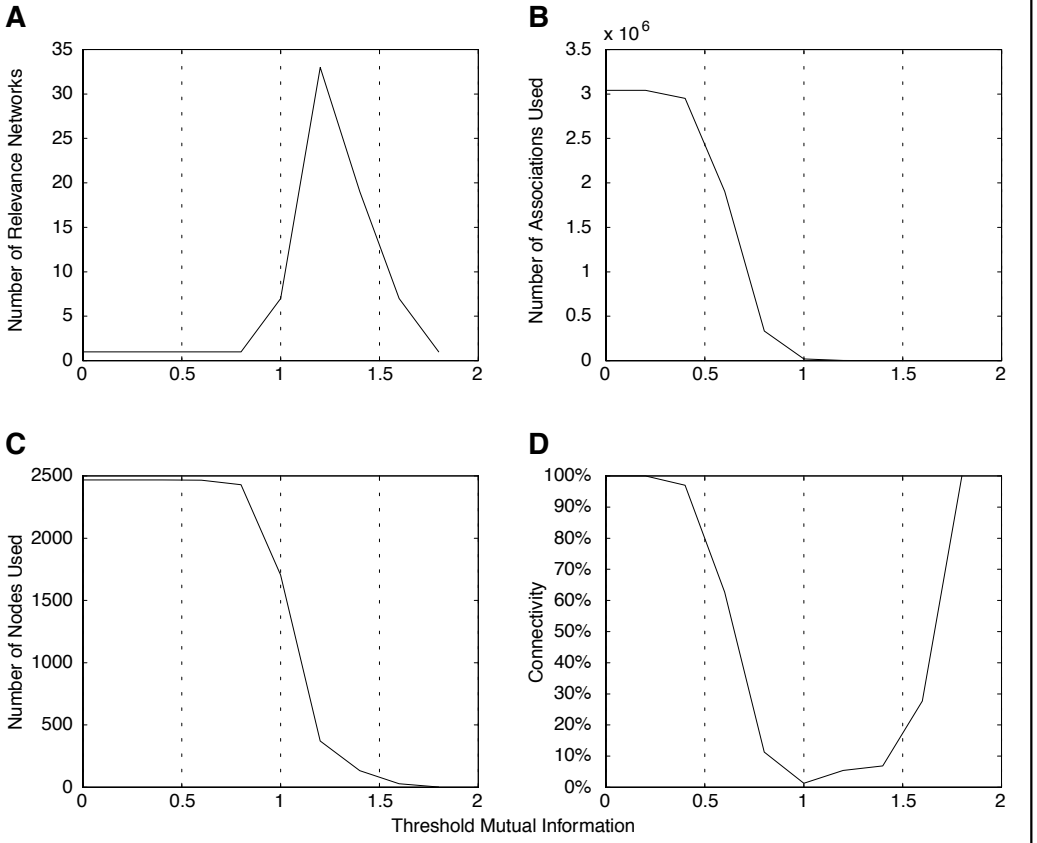
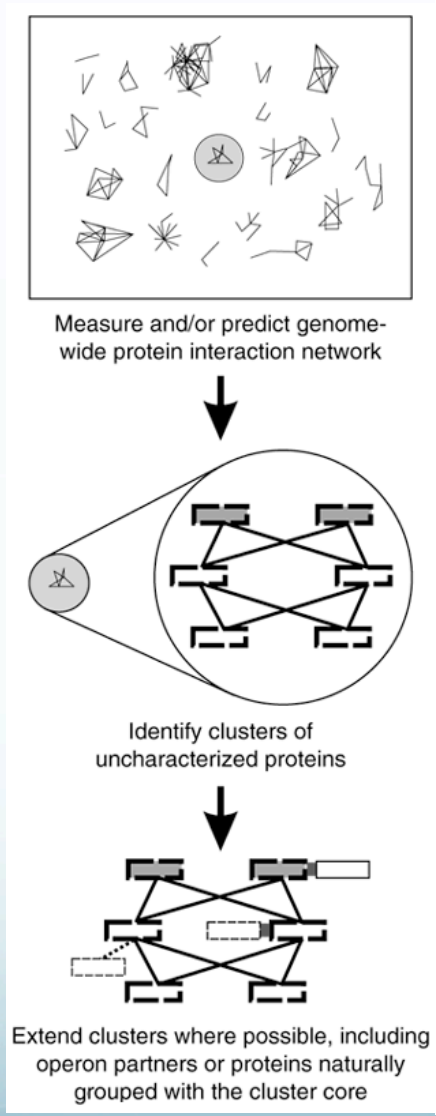


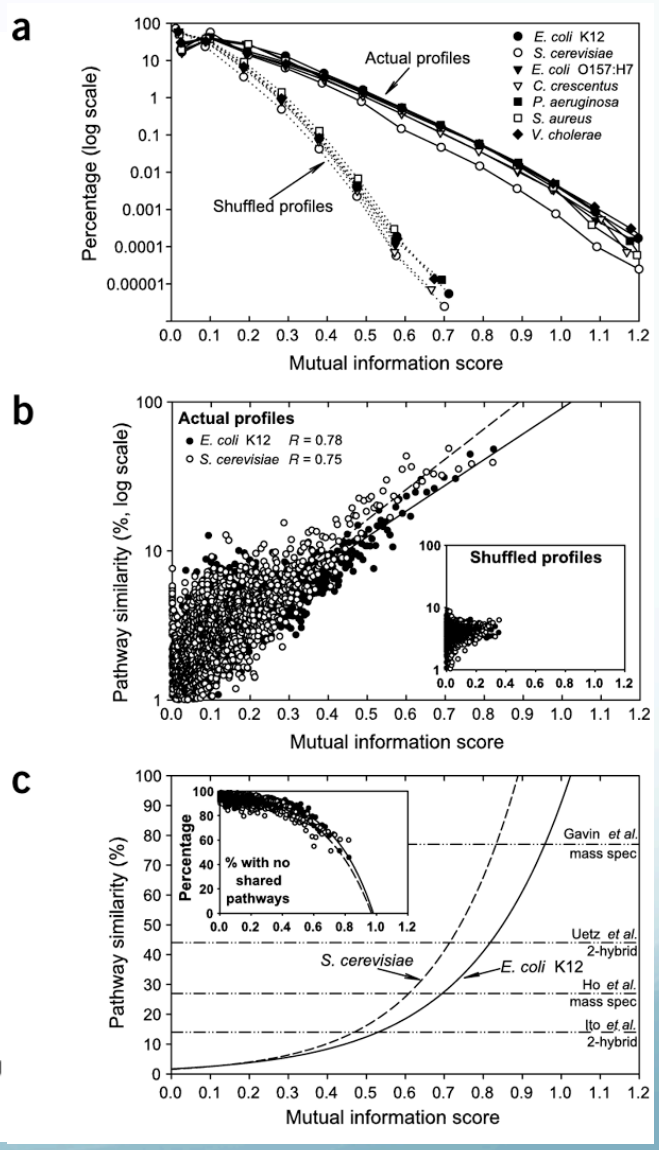
Figure 2: (A) Number of Relevance Networks as a function of the chosen threshold mutual information (TMI). (B) Number of generated hypothetical gene-gene associations versus TMI. (C) Number of genes used versus TMI. (D) Connectivity of Relevance Networks versus TMI.

Butte and Kohane. 2000. Pacific Symposium in Biocomputing. 418-429.

MI – Discovering functional linkages

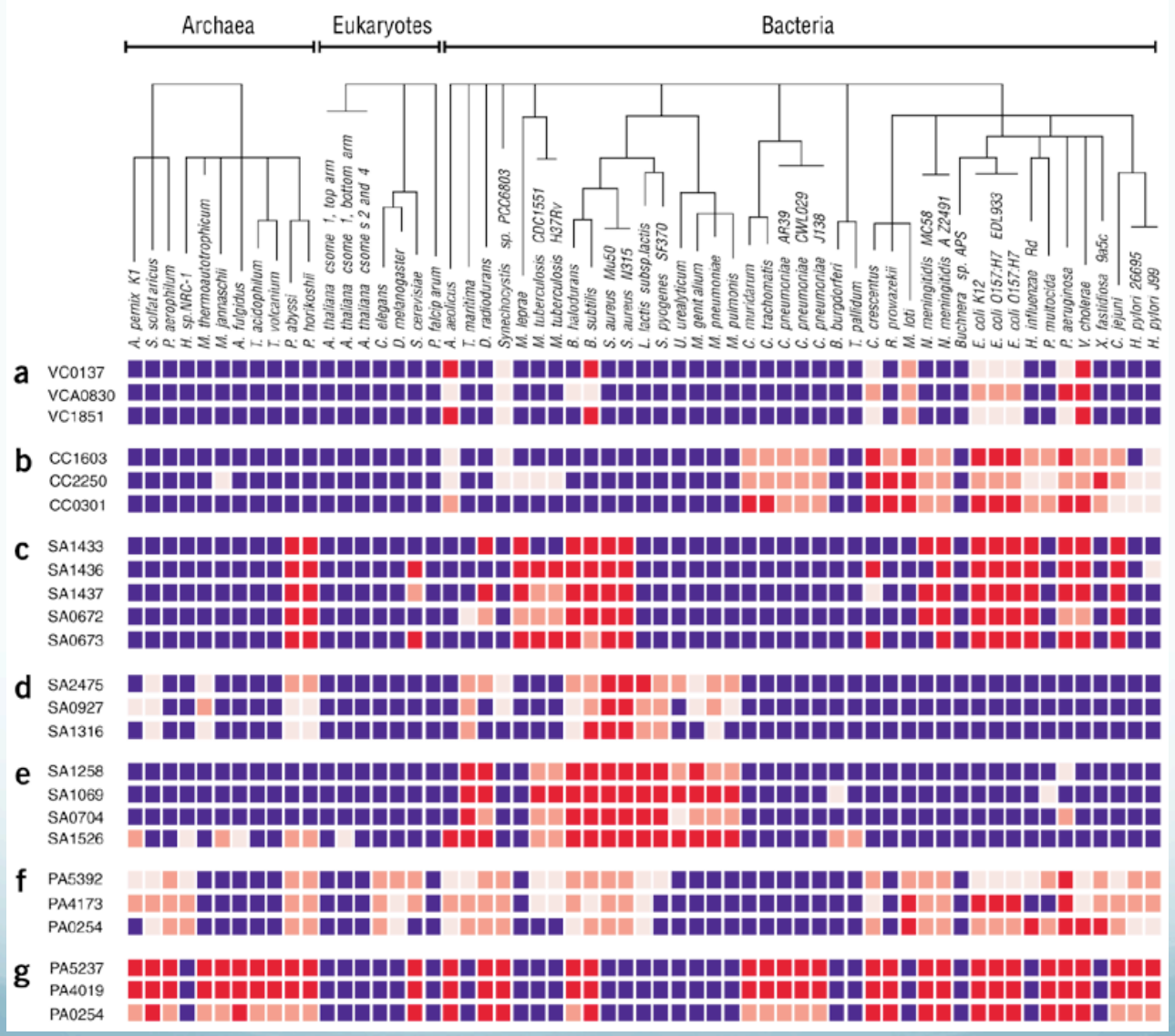
- Used phylogenetic profiles of proteins
- Instead of each element being 0 or 1, each element was -1/log(E-value)
- Used MI as a metric to cluster proteins based on their profiles





Date and Marcotte. 2003. Nature Biotechnology 21(9), 1055-1062.

MI – Discovering functional linkages



Date and Marcotte. 2003. Nature Biotechnology 21(9), 1055-1062.

Other applications of MI

- Bindewald and Shapiro (2006) used MI between positions on sequence alignments as a feature for the **prediction of RNA secondary structure**
- Tomovic and Oakeley (2007) used MI in **transcription factor binding site analysis** to identify highly correlated positions
- Buslje et al. (2010) have shown that **networks of high MI define the structural proximity of catalytic sites** and can be used for their prediction
- Brunel et al. (2010) have devised a 'mutual information statistical significance' test for **genetic association studies**

Thanks for listening