Expression Bioinformatics: Peer Teaching – Assignment 8

Stefan Altendorfer, Annika Ladwig, Kilian Hunter

Information Entropy

Information entropy

- measures uncertainty/variance associated with a variable
  - No uncertainty → no information content, since outcome is always the same

$$H = -\sum p_i \log_2(p_i)$$

Joint information entropy of X and Y:

- measures uncertainty associated with two variables
- Refers to the joint probability

$$H(\mathbf{x}, \mathbf{y}) = -\sum_{i=1}^{n} p(x_i, y_i) \log_2(p(x_i, y_i)$$

Why is $H(x,y) \leq H(x) + H(y)$?

- X and Y are independent
  - x does not provide any information about the outcome of y (vice versa)
  - $H(x,y) = H(x) + H(y)$

- X and Y are NOT independent
  - some degree of correlation between x and y → x provides information about outcome of y (depends on level of correlation)
  - $H(x,y) < H(x) + H(y)$

➢ Joint entropy is always less than or equal to the sum of individiual entropies

➢ 2 variables cannot increase uncertainty beyond what is present when considering them separately

It was said the M(x,y) to be 0 when the samples are independent. But then we divide it in the normalization. How does it come?

- Mutual Information
  - Quantifies the amount of information gained about one variable by knowing the value of another.
  - Reflects the statistical dependence or association between variables
- M(x,y) = H(x) + H(y) – H(x,y) = H(x) + H(y) – H(x) – H(y) = 0
- Max-Scaling: [0,1]
- 1 ≙ max(H(x),H(y))
- For M(x,y) = 0 technically no further scaling is needed, but does not hurt (should be applied to each value of the distribution)

$$M(x,y)_{norm} = \frac{M(x,y)}{\max\left(H(x), H(y)\right)}$$

Question 3: Wth is this graph?

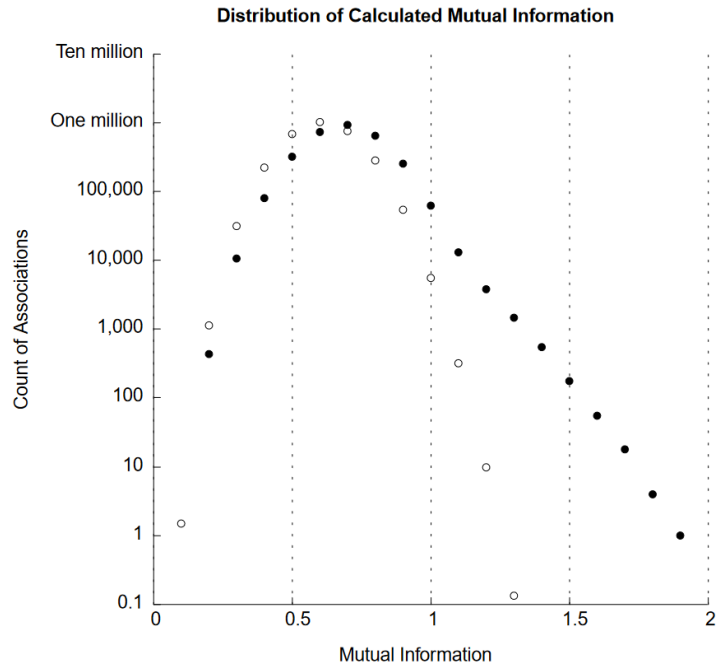


Figure 1: Pair-wise mutual information was calculated between 79 measurements of RNA expression of 2,467 genes in *Saccharomyces cerevisiae* and the distribution of these is shown with filled circles. The same was calculated using permuted RNA expression measurements; the average distribution from 30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

Question 3: Wth is this graph?

- Aim was to find functional genomic cluster
- Computation of pairwise MI values for 2,467 genes with 79 expression measurements (black dots)
- Permutation of measurements, then computation of MI values (white dots)
- Count of Associations: how many genes are associated with this MI value
- White dots represent MI values that occur by random chance -> black dots outside that range represent significant associations (here: MI > 1.3)

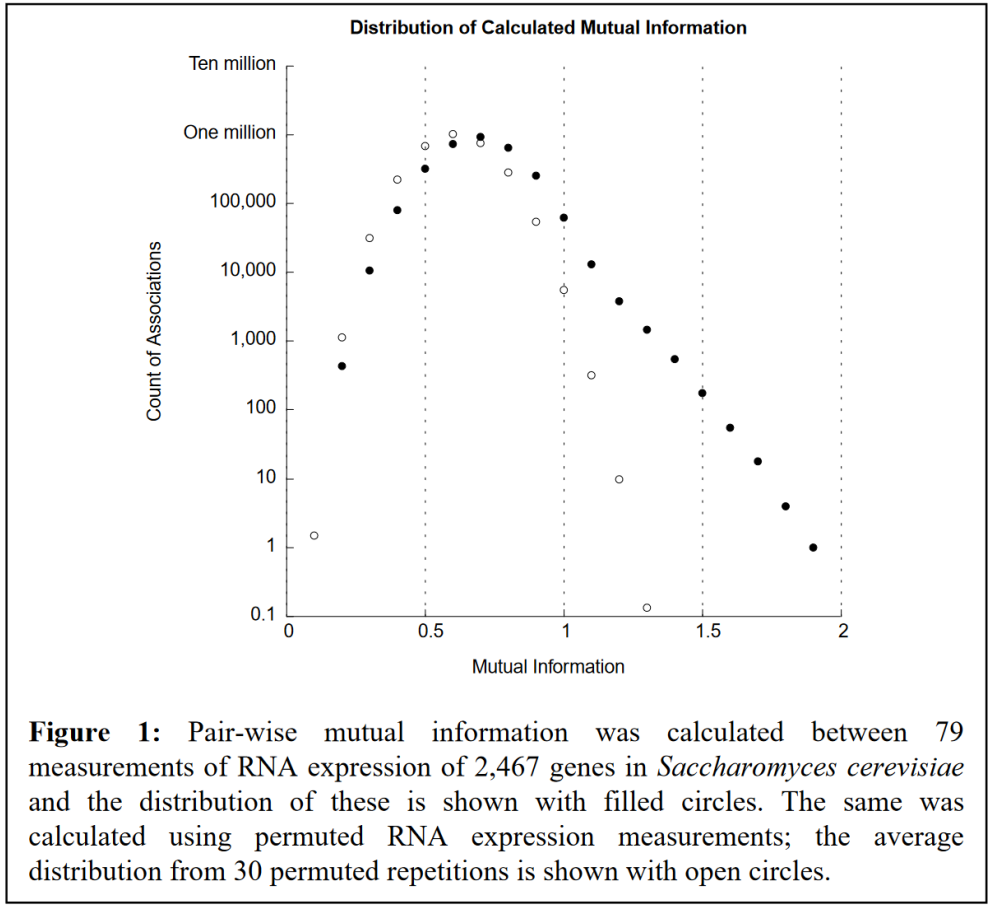


Figure 1: Pair-wise mutual information was calculated between 79 measurements of RNA expression of 2,467 genes in *Saccharomyces cerevisiae* and the distribution of these is shown with filled circles. The same was calculated using permuted RNA expression measurements; the average distribution from 30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

MI- How to:

	0h	1h	2h	3h	4h	5h	6h	7h	8h	9h
G1	8.0	7.6	7.1	6.3	6.0	6.1	5.3	5.1	4.5	4.0
G2	5.0	4.7	5.1	5.6	5.0	4.9	5.6	5.9	5.4	5.2

MI- How to:

	0h	1h	2h	3h	4h	5h	6h	7h	8h	9h
G1	8.0	7.6	7.1	6.3	6.0	6.1	5.3	5.1	4.5	4.0
G2	5.0	4.7	5.1	5.6	5.0	4.9	5.6	5.9	5.4	5.2

1. Discretization
2. Estimate probabilities of possible outcomes for each variable
3. Compute individual entropies for each variable → $H = -\sum p_i \log_2(p_i)$
4. Compute Joint Probabilities
5. Compute Joint Information Entropy → $H(\mathbf{x}, \mathbf{y}) = -\sum_{i=1}^{n} p(x_i, y_i) \log_2(p(x_i, y_i)$
6. Compute Mutual Information Content → $M(x, y) = H(x) + H(y) - H(x, y)$
7. Normalize Mutual Information Content → $M(x, y)_{norm} = \frac{M(x, y)}{\max(H(x), H(y))}$
8. Transform into Distance → $d_{MI}(x, y) = 1 - M(x, y)_{norm}$

Mutual Information Example

	0h	1h	2h	3h	4h	5h	6h	7h	8h	9h
Gene 1	8.0	7.6	7.1	6.3	6.0	6.1	5.3	5.1	4.5	4.0
Gene 2	5.0	4.7	5.1	5.6	5.0	4.9	5.6	5.9	5.4	5.2

How to discretize time series data in a meaningful way?

Solution

1. Discretization:

	0h	1h	2h	3h	4h	5h	6h	7h	8h	9h
Gene 1	0	0	0	0	0	0	0	0	0	0
Gene 2	0	0	1	1	0	0	1	1	0	0

2. Estimate probabilities:

For gene 1: $p(0) = 1$, $p(1) = 0$

For gene 2: $p(0) = 0.6$, $p(1) = 0.4$

3. Individual entropies:

$H(g1) = -(10 * 1 * \log2(1)) = 0$

$H(g2) = -(6 * 0.6 * \log2(0.6) + 4 * 0.4 * \log2(0.4)) \approx 4.77$

Solution

4. Joint probabilities:

$P(0, 0) = 0.6$

$P(1, 1) = 0$

$P(0, 1) = 0.4$

$P(1, 0) = 0$

5. Joint mutual information entropy:

$H(g1, g2) = 6 * 0.6 * \log2(0.6) + 4 * 0.4 * \log2(0.4) \approx 4.77$

6. Mutual information content:

$MI(g1, g2) = H(g1) + H(g2) - H(g1, g2) = 0 + 4.77 - 4.77 = 0$

Solution

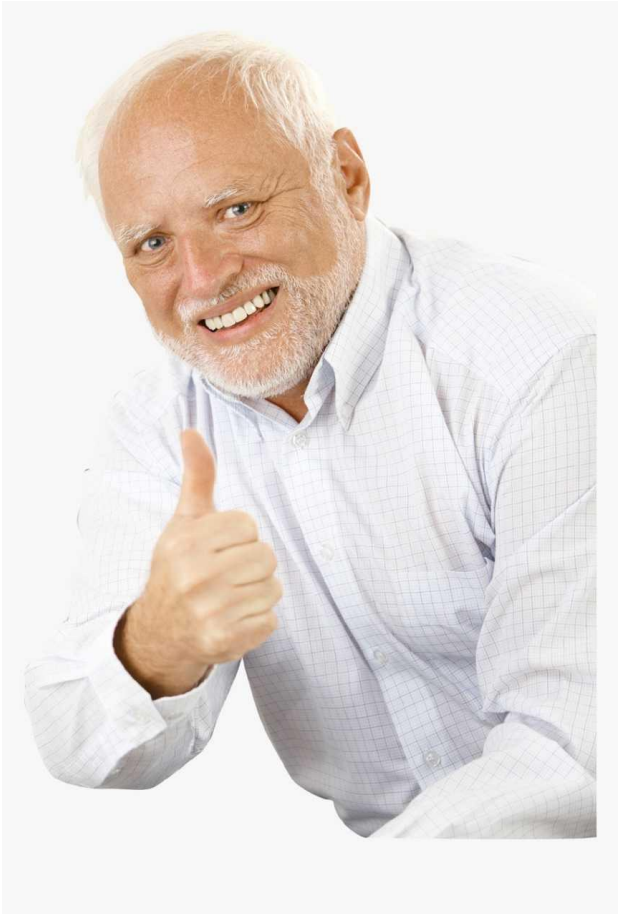
7. Normalize mutual information content:

$MI\_norm = MI(g1, g2)/\max(H(g1), H(g2)) = 0/4.77 = 0$

8. Transformation into distance:

$d\_MI(g1, g2) = 1 - MI\_norm = 1 - 0 = 1$

Questions?



Next:

t1p.de/expbiopt1

Questions

In data analysis, when might you prefer to use Euclidean distance over Spearman distance?

A) When dealing with categorical variables
B) When the data exhibits monotonic relationships
C) When outliers have a significant impact on similarity measures
D) When you want to measure the angular separation between vectors

What is the primary purpose of introducing random perturbation in a system or dataset?

A) To eliminate all sources of variability
B) To introduce controlled random variations or disturbances
C) To determine deterministic outcomes
D) To increase precision in measurements