*FULL PAPER*

# Lattices of Random Variables

**Zhongyuan Che**

Penn State Beaver, 100 University Drive, Monaca, PA 15061
Phone: 724-773-3884, Fax: 724-773-3557, Email: *zxc10@psu.edu*

**Michael D. Rice**

Department of Mathematics and Computer Science, Wesleyan University, Middletown, CT 06459
Phone: 860-685-2184, Fax: 860-685-2571, Email: *mrice@wesleyan.edu*

**Abstract** We provide a rigorous proof of Shannon's assertion that the set of random variables on a fixed sample space has a natural partial order defined in terms of conditional entropy and furthermore, is a lattice in the mathematical sense. Moreover, we show that in a restricted setting the lattice is isomorphic to the lattice of partitions of a set. We also study relationships between the partial order and mutual information that provide a natural connection with the analysis of alignments of RNA sequences.

**Keywords** partial order, lattice, random variable, entropy, mutual information, sequence analysis

**AMS Mathematics Subject Classification Codes**
**Primary** 94A17, 06B05
**Secondary** 05C05

## Introduction

In the short paper [10], Claude Shannon defined a special relation on random variables in terms of conditional entropy. He claimed that the relation was a partial order on random variables and, in fact, that it was also a lattice (in the mathematical sense [6]). Apparently, other authors have not studied this idea. Recently, we discovered Shannon's paper and realized that the order notion was potentially useful for our work in bioinformatics. In the course of our work, we found it necessary to further develop the idea. The present paper summarizes the theoretical aspects of our investigations. A future paper [4] will describe how the ideas are useful for developing a general model of comparative sequence analysis.

We feel that the present paper makes two principal contributions. First, it provides a rigorous foundation for the order-theoretic and lattice ideas introduced by Shannon. Secondly, the paper reintroduces the order-theoretic ideas to the general community and connects these ideas with other familiar ideas about random variables.

## Organization

In Section 1, we investigate the order relation $\leq$ defined by Shannon in [10] and provide rigorous proofs of his assertions. The main tool used is Proposition 1.1 that shows the order notion is equivalent to a special functional relationship between random variables. Proposition 1.3 establishes that $\leq$ is a partial order (henceforth simply called an order) and Proposition 1.7 establishes that the ordered set is a lattice. In this context, we note that the least upper bound of two random variables corresponds to the familiar notion of a joint random variable while the greatest lower bound does not seem to have a natural interpretation.

In Section 2, we study the order relation in a slightly more restricted setting. In this setting, Proposition 2.1 establishes that the order relation on random variables is equivalent to the refinement relation on certain partitions associated with the random variables. Based on this fact, Proposition 2.2 establishes that Shannon's lattice is isomorphic to a special lattice of partitions. Section 2 also presents several examples of the order relation based on alignments of RNA sequences.

In Section 3, we investigate the relationships between the order relation and mutual information. In Proposition 3.1, we observe that the mutual information between two random variables is maximized precisely when the random variables are order-related. In Proposition 3.3, we establish that the uncertainty of the greatest lower bound of two random variables is a lower bound on the mutual information. In general, this lower bound is a strict inequality, but in Proposition 3.4 we prove that equality holds assuming a special relationship between the random variables. The relationship is a rather technical graph-theoretic condition, but it provides a starting point for further investigations.

The ideas presented in the paper have many evident connections with ideas that are well known in information theory as well as with recent work in the field of bioinformatics. These connections are discussed in Section 4.

The paper contains two appendices. Appendix 1 contains a brief summary of the technical terms and standard results that are used in the proofs. Appendix 2 contains the details of the longer proofs. This technique has been adopted to avoid confusing the key ideas and results in the main text.

## Section 1 – A Partial Order on Random Variables

In [10], Shannon presented the following definition of an order $\leq$ between random variables.

**Definition 1.1** Given random variables $X$ and $Y$, define $X \leq Y$ if $H(X \mid Y) = 0$.

The definition of $X \leq Y$ says essentially that the values of $X$ are determined by the values of $Y$. The precise relationship is stated in the following result.

**Proposition 1.1** The following statements are equivalent for random variables $X$ and $Y$:

(a) $X \leq Y$.

(b) For each value $y_j$ of $Y$, there is a unique value $x_i$ of $X$ such that $p(x_i \mid y_j) = 1$.

(c) There exists a surjection $\varphi : \{y_1,\ldots, y_m\} \rightarrow \{x_1,\ldots, x_n\}$ such that

$$P(X = \varphi(y_j) \wedge Y = y_j) = P(Y = y_j).$$

(d) There exists a surjection $\varphi : \{y_1,\ldots, y_m\} \rightarrow \{x_1,\ldots, x_n\}$ such that $P(X = \varphi \circ Y) = 1$.

It follows from Proposition 1.1 that if $X \leq Y$, then $|\, Range(X) \,| \leq |\, Range(Y) \,|$.

**Definition 1.2** Given random variables $X$ and $Y$, define $X \equiv Y$ if $X \leq Y$ and $Y \leq X$.

It follows from Proposition 1.1 that $X \equiv Y$ if and only if $X \leq Y$ and $|\, Range(X) \,| = |\, Range(Y) \,|$. (If $X \leq Y$ and $|\, Range(X) \,| = |\, Range(Y) \,|$, then the surjection $\varphi$ in part (d) is a bijection. Therefore, $P(Y = \varphi^{-1} \circ X) = P(X = \varphi \circ Y) = 1$, so $Y \leq X$.)

The following two results can be derived based on the above definition and Proposition 1.1.

**Proposition 1.2** The relation $\equiv$ is an equivalence relation on the set of random variables.

**Proposition 1.3** The relation $\leq$ is a partial order on the equivalence classes $L$ (with respect to $\equiv$) and the entropy function $H$ is an order-preserving mapping on $L$.

Let $\perp$ denote the equivalence class of random variables $[X]$ where each $X$ has a single value $x$ that satisfies $p(x) = 1$. Then $\perp$ is the smallest element in $L$ since by Proposition 1.1, $\perp \leq Y$ for any random variable $Y$. Therefore, it follows from the remark after Definition 1.2 that each finite chain $C \subseteq L$ contains a random variable $X$ that satisfies $|\, Range(X) \,| \geq |\, C \,|$.

We will now establish that the partially ordered set $L$ is, in fact, a lattice. For simplicity, we will work with random variables instead of equivalence classes of random variables.

**Proposition 1.4** Given random variables $X$ and $Y$, $Z = (X, Y)$ is the *smallest* random variable satisfying $X \leq Z$ and $Y \leq Z$. This random variable will be denoted $X \vee Y$.

The construction of the *largest* random variable $Z = X \wedge Y$ that satisfies $Z \leq X$ and $Z \leq Y$ is somewhat more complicated and requires the following preliminary result.

**Proposition 1.5** Given random variables $X$ and $Y$, $X \le Y$ if and only if there exists a mapping $\lambda : \{y_1,\ldots,y_m\} \to \{y_1,\ldots,y_m\}$ such that $X \equiv \lambda \circ Y$. Hence, up to equivalence, there exist only *finitely* many random variables $X$ that satisfy $X \le Y$.

The following result confirms the existence of the meet $X \wedge Y$ of two random variables.

**Proposition 1.6** Given random variables $X$ and $Y$, $X \wedge Y = \vee \mathbf{F}$, where the family $\mathbf{F} = \{A_r\}$ consists of one random variable from each equivalence class $[A_r]$ such that $A_r$ satisfies $A_r \le X$ and $A_r \le Y$.

**Proposition 1.7** Based on the operations $\vee$ and $\wedge$, the partially ordered set $L$ is a lattice.

To establish Proposition 1.7, first note that by Proposition 1.3, $(L, \le)$ is a partially ordered set. Also, Propositions 1.4 and 1.6 show that equivalence classes $[X]$ and $[Y]$ have a smallest upper bound $[X \vee Y]$ and a greatest lower bound $[X \wedge Y]$. Therefore, by definition, $L$ is a lattice ([6], 2.4(i)). A direct construction of $X \wedge Y$ will be presented in Section 2 for the special case where $P(\{s\}) > 0$ for each event $s \in S$.

The lattice operations $\vee$ and $\wedge$ can be characterized in the following manner:

(a) The join $X \vee Y$ is the unique random variable (up to equivalence) that satisfies $X \le X \vee Y$ and $Y \le X \vee Y$ and has uncertainty $H(X, Y)$.

(b) The meet $X \wedge Y$ is the unique random variable (up to equivalence) that satisfies $X \wedge Y \le X$ and $X \wedge Y \le Y$ and has uncertainty $H(X \wedge Y)$.

To establish part (a), suppose $P$ is a random variable that satisfies $X \le P$, $Y \le P$, and $H(P) = H(X, Y)$. By Proposition 1.4, $H(P) = H(X \vee Y)$ and $X \vee Y \le P$, so $H(X \vee Y \mid P) = 0$. Also, by equation (4) in Appendix 1,

$$\begin{aligned} H(P \mid X \vee Y) &= H(X \vee Y, P) - H(X \vee Y) \\ &= H(P) + H(X \vee Y \mid P) - H(X \vee Y) \\ &= H(P) - H(X \vee Y) = 0 \end{aligned}$$

so $P \le X \vee Y$. Therefore, it follows that $P \equiv X \vee Y$. A similar proof establishes part (b).

As an example of how the lattice-theoretic point of view facilitates arguments, we will show that for each random variable $X$, the conditional function $H(\bullet \mid X)$ is an order-preserving mapping on $L$. (A slightly restricted version of this statement is given in Theorem 2.8 of [12].)

By equation (4) in Appendix 1, $H(A \mid X) = H(A, X) - H(X)$ and $H(B \mid X) = H(B, X) - H(X)$. Therefore, $H(A \mid X) - H(B \mid X) = H(A, X) - H(B, X) = H(A \vee X) - H(B \vee X)$. If $A \le B$, then $A \vee X \le B \vee X$; hence by Proposition 1.3, $H(A \vee X) - H(B \vee X) \le 0$, so $H(A \mid X) \le H(B \mid X)$.

## Section 2 - Partition Lattices

In this section, we will show that under an additional restriction, the lattice $L$ discussed in Section 1 is essentially the partition lattice that is studied in the field of combinatorics.

The key restriction is that the probability measure $P$ satisfies the following condition: $P(\{s\}) > 0$ for each event $s \in S$. This assumption guarantees that $p(x) > 0$ for each value $x$ of a random variable defined on $S$ so the results in Section 1 are still valid.

With this additional restriction on the probability space $(S, P)$, one can establish a strengthened version of Proposition 1.1. This requires the following definition.

**Definition 2.1** (a) Given a random variable $X$, define the partition $\Gamma_X = \{X^{-1}(x_i) \mid 1 \le i \le m\}$ of $S$.
(b) Given partitions $\Gamma$ and $\Phi$ of $S$, define $\Gamma < \Phi$ ($\Gamma$ *refines* $\Phi$) if for each $U \in \Gamma$, there exists $V \in \Phi$ such that $U \subseteq V$.

The following result is an analogue of Proposition 1.1 that replaces the equivalence relation $\equiv$ with equality = and shows that the ordering $\le$ is set-theoretic.

**Proposition 2.1** The following statements are equivalent for random variables $X$ and $Y$:
(a) $X \le Y$.
(b) There exists a surjection $\varphi : \{y_1,\ldots, y_m\} \to \{x_1,\ldots, x_n\}$ such that $X = \varphi \circ Y$.
(c) $\Gamma_Y < \Gamma_X$.

Based on Definition 1.1, the equivalence of (a) and (c) is also established in Theorem 2.3 of [12].

It follows from Proposition 2.1 that in our more restricted setting, the notion of equivalence corresponds to equality of partitions: $X \equiv Y$ if and only if $\Gamma_X = \Gamma_Y$. (Based on Definition 1.2 and Proposition 2.1, $X \equiv Y$ is equivalent to $\Gamma_X < \Gamma_Y$ and $\Gamma_Y < \Gamma_X$ and this occurs precisely when $\Gamma_X = \Gamma_Y$.)

The standard *partition lattice* $L(S)$ on a set $S$ consists of all partitions of $S$, where the partial order is the refinement operation $<$, the greatest lower bound of two partitions $\Gamma$ and $\Phi$ is the partition $\Gamma \wedge \Phi = \{A \cap B \mid A \in \Gamma, B \in \Phi\}$ , and the least upper bound of $\Gamma$ and $\Phi$ is the partition $\Gamma \vee \Phi$ consisting of the intersection of all partitions that are refined by both $\Gamma$ and $\Phi$ ([2], Example 9, p. 15; a complete discussion is given in [9]).

For the current work, it is convenient to reverse the ordering on $L(S)$, that is, to define the partial order $\le_*$ by $\Phi \le_* \Gamma$ if $\Gamma < \Phi$. Then $(L(S), \le_*)$ is still a lattice, but the roles of $\Gamma \wedge \Phi$ and $\Gamma \vee \Phi$ are interchanged. The following result shows that this lattice is equivalent to the lattice introduced in Section 1.

**Proposition 2.2** The lattices $(L,\le)$ and $(L(S), \le_*)$ are order-isomorphic and lattice-isomorphic.

Without presenting the complete details of the proof (found in Appendix 2), an order-isomorphism from $L$ to $L(S)$ is defined by the mapping $[X] \to \Gamma_X$ that associates the partition $\Gamma_X$ with an equivalence class $[X]$ of random variables. Also, it follows from general considerations that an order-isomorphism between two lattices is also a lattice isomorphism ([6], 2.19(ii)). In particular, for random variables $X$ and $Y$, $X \vee Y$ corresponds to $\Gamma_X \wedge \Gamma_Y$ and $X \wedge Y$ corresponds to $\Gamma_X \vee \Gamma_Y$.

It follows from Proposition 2.2 that the meet $X \wedge Y$ of random variables $X$ and $Y$ can be constructed in the following manner. Let $\Gamma$ denote the intersection of all partitions $\Phi$ of $S$ that satisfy $\Gamma_X < \Phi$ and $\Gamma_Y < \Phi$. Clearly, $\Gamma = \{A_k\}$ is the *smallest* partition of $S$ such that $\Gamma_X < \Gamma$ and $\Gamma_Y < \Gamma$. Then the random variable $X \wedge Y : S \to \{z_k\}$ is defined by $(X \wedge Y)(s) = z_k$ if $s \in A_k$. However, a direct construction of $X \wedge Y$ (or equivalently $\Gamma_X \vee \Gamma_Y$) can be given by using the following graph $wpg(X, Y)$.

**Definition 2.2** Given random variables $X$ and $Y$, suppose $\Gamma_X = \{A_i\}$ and $\Gamma_Y = \{B_j\}$. The vertex set of $wpg(X, Y)$ is the disjoint union $\{A_i\} \cup \{B_j\}$ of the two partitions and the edge set consists of the pairs $(A_i, B_j)$ such that $A_i \cap B_j \neq \emptyset$.

**Proposition 2.3** Given random variables $X$ and $Y$ defined on $S$,

$$\Gamma = \{\cup C \mid C \text{ is a } \textit{connected component} \text{ of } wpg(X, Y)\}$$

is the *smallest* partition of $S$ such that $\Gamma_X < \Gamma$ and $\Gamma_Y < \Gamma$; hence $\Gamma = \Gamma_X \vee \Gamma_Y$.

Using Propositions 2.1 - 2.3, it is straightforward to analyze certain partially ordered sets of random variables. An important example of this type is found in the field of bioinformatics. Assume that $S = \{s_1, s_2, \ldots, s_m\}$ is a set of RNA sequences each of length $n$. Then each position $1 \leq j \leq n$ determines a random variable

$$X_j : S \to \{A, C, G, U\}$$

that assigns to each sequence $s_i$ the nucleotide $s_{ij}$ at position $j$. Assume that the probability $p_j(\alpha) = P(X_j = \alpha)$ denotes the *frequency of occurrence* of the nucleotide $\alpha$ in position $j$ and the probability $p_{jk}(\alpha, \beta)$ denotes the frequency of occurrence of the di-nucleotide pair $\alpha\beta$ in positions $j$ and $k$, respectively. Therefore, the assumption made in this section is satisfied ($P(\{s\}) > 0$ for each $s \in S$).

Let $P$ denote the partially ordered set of random variables $\{X_j \mid 1 \leq j \leq n\}$ and let $L$ denote the *smallest* lattice of random variables that contains $P$. Based on Proposition 1.7, the lattice $L$ exists. Figure 1 illustrates the notion of order in this context. (For convenience, we will refer to the sequences by number.)

| $S$ | $X_1$ | $X_2$ | $X_3$ | ... |
|---|---|---|---|---|
| 1 | U | G | A | ... |
| 2 | U | A | C | ... |
| 3 | A | U | C | ... |
| 4 | A | U | C | ... |

**Figure 1 – Order in Sequence Alignment**

By Proposition 2.1, $X_1 \leq X_2$ since $\Gamma_{X_2} = \{\{1\}, \{2\}, \{3, 4\}\} < \Gamma_{X_1} = \{\{1, 2\}, \{3, 4\}\}$. On the other hand, $X_2 \leq X_1$ does not hold since $\Gamma_{X_1} < \Gamma_{X_2}$ fails. Since $\Gamma_{X_3} = \{\{1\}, \{2, 3, 4\}\}$, it follows from Proposition 2.2 that $X_1 \vee X_3$ is represented by the partition $\{\{1\}, \{2\}, \{3, 4\}\}$. Also, $X_1 \wedge X_3 \equiv \perp$ since it follows from Proposition 2.3 that $\Gamma_{X_1} \vee \Gamma_{X_3} = \{S\}$.

A more realistic example of order is based on the alignment shown in Figure 2. This alignment represents positions 1 – 17 in a set of *tRNA* sequences (of length 97) found in the Bayreuth database ([11]).

| Organism | Sequence |
|---|---|
| RA1140 UGC MYCOPLASMA CAPRIC. EUBACT | GGGCCCUUAGCUCAGCU |
| RA1180 UGC MYCOPLASMA MYCOID. EUBACT | GGGCCCUUAGCUCAGCU |
| RA1540 5GC BACILLUS SUBTILIS EUBACT | GGAGCCUUAGCUCAGCU |
| RA1660 GGC E.COLI EUBACT | GGGGCUAUAGCUCAGCU |
| RA1661 VGC E.COLI EUBACT | GGGGGCAUAGCUCAGCU |
| RA1662 VGC E.COLI EUBACT | GGGGCUAUAGCUCAGCU |
| RA3920 UGC NEUROSPORA CRASSA MI SIN | GGGGGUAUAGUAUAAUU |
| RA7630 IGC SACCHAROMYCES CER. CY SIN | GGGCGUGUGGCGUAGUC |
| RA7650 IGC TORULOPSIS UTILIS CY SIN | GGGCGUGUGGCGUAGUU |
| RA9230 IGC BOMBYX MORI CY ANI | GGGGGCGUAGCUCAGAU |

**Figure 2 – *tRNA* Sequence Alignment**

There is complete conservation at positions 1, 2, 8, 10, and 14, so the equivalence class $\{X_1, X_2, X_8, X_{10}, X_{14}\}$ represents the bottom element $\perp$. In addition, $X_{11} \equiv X_{15}$, but neither random variable is equivalent to $\perp$. Up to equivalence, the positions $(i, j)$ that satisfy $X_i \leq X_j$, where $X_i$ is not equivalent to $\perp$, are (9,12), (11,12), (13,12), and (13,16).=

Of course, the set $P$ is not the entire lattice $L$. For example, $X_{12} \vee X_{16}$ is represented by the partition {{1-6}, {7}, {8,9}, {10}} so it is not equivalent to any $X_i$. On the other hand, $X_{12} \wedge X_{16}$ is represented by the partition {{1-6,10}, {7-9}}, so $X_{12} \wedge X_{16} \equiv X_{13}$.

## Section 3 – Mutual Information

In this section, we will assume the restriction that was introduced in Section 2. Given random variables $X$ and $Y$, the *mutual information* between $X$ and $Y$ can be interpreted as a measure of the *covariation* of $X$ and $Y$, that is, the degree to which changes in the values of one random variable are reflected by changes in the other. In particular, $X$ and $Y$ are independent if and only if $M(X, Y) = 0$.

Based on equation (6) in Appendix 1, the mutual information between $X$ and $Y$ is

$$M(X, Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X).$$

Since each value $H(\bullet \mid \bullet) \geq 0$, it follows that $M(X, Y) \leq \mathbf{min}\{H(X), H(Y)\}$ and the minimum is achieved precisely when at least one of the conditional entropies is zero. Therefore, the following result is an immediate consequence of Definitions 1.1 and 1.2.

**Proposition 3.1** Given random variables $X$ and $Y$,

(a) $X \leq Y$ if and only if $M(X, Y) = H(X)$.

(b) $X \equiv Y$ if and only if $M(X, Y) = H(X) = H(Y)$.

As an illustration, in Figure 1, $M(X_1, X_2) = H(X_1) = 1$ since $X_1 \leq X_2$ while $M(X_1, X_2) < H(X_2) = 1.5$.

The equality $M(X, Y) = H(X)$ has also been noted by other authors without their introducing the partial order $\leq$ that is the focus of the current paper. Several authors have used mutual information to predict a common secondary structure from an alignment of RNA sequences. This approach, known as comparative sequence analysis, is based on computing the co-variation between pairs of positions in the sequences. A survey of the technique is found in [8].

The idea of using mutual information to predict secondary structure was introduced in [3]. In essence, the approach computes positions $i$ and $j$ in the sequences where the mutual information $M(X_i, X_j)$ is *large*. Since $M(X_i, X_j) \leq \mathbf{min}\{H(X_i), H(X_j)\}$ and $H(X_i) \leq 2$ (assuming no gaps at position $i$), this approach tends to identify positions where an ordering between $X_i$ and $X_j$ approximately holds. A study of comparative sequence analysis in terms of the partial order discussed in this paper will be presented in [4].

The next result establishes an analogue of Proposition 1.3 for mutual information.

**Proposition 3.2** For each random variable $X$, the mutual information function $M(\bullet, X)$ is an order-preserving mapping on $L$.

It follows from Proposition 3.1 that if $X \leq Y$, then $M(X, Y) = H(X) = H(X \wedge Y)$. Therefore, it is natural to enquire about a general relationship between $H(X \wedge Y)$ and $M(X, Y)$. The next result shows that the uncertainty of the meet of two random variables is a lower bound on their mutual information.

**Proposition 3.3** For each pair of random variables $X$ and $Y$, $H(X \wedge Y) \leq M(X, Y)$.

To establish Proposition 3.3, note that by definition $X \wedge Y \leq Y$, so it follows from Proposition 3.1 that $H(X \wedge Y) = M(X \wedge Y, Y)$. Also, it follows from Proposition 3.2 that $M(X \wedge Y, Y) \leq M(X, Y)$. Therefore, $H(X \wedge Y) \leq M(X, Y)$.

It follows from Proposition 3.3 that if $M(X, Y) = 0$, then $H(X \wedge Y) = 0$. However, in general, the equality $H(X \wedge Y) = M(X, Y)$ does not hold. For example, in Figure 1, $X_1 \wedge X_3 \equiv \perp$, so $H(X_1 \wedge X_3) = 0$, but $M(X_1, X_3) = 3(2 - \log_2(3))/4 \approx 0.31$.

Notice that based on Figure 1, the partition graph $wpg(X_1, X_3)$ contains the following path of length 3:

$$X_3^{-1}(\{\mathrm{A}\}) — X_1^{-1}(\{\mathrm{U}\}) — X_3^{-1}(\{\mathrm{C}\}) — X_1^{-1}(\{\mathrm{A}\}).$$

(For example, $X_3^{-1}(\{\mathrm{A}\}) = \{1\}$ and $X_1^{-1}(\{\mathrm{U}\}) = \{1, 2\}$, so $X_3^{-1}(\{\mathrm{A}\}) \cap X_1^{-1}(\{\mathrm{U}\}) \neq \emptyset$, …)

However, there does not exist a path of length 3 in $wpg(X, Y)$ if $X \leq Y$. (Otherwise, there exist distinct members $U, U' \in \Gamma_X$ and $V, V' \in \Gamma_Y$ such that $U \cap V \neq \emptyset$, $U' \cap V' \neq \emptyset$, and either (a) $U' \cap V \neq \emptyset$ or (b) $U \cap V' \neq \emptyset$. Since $X \leq Y$, in case (a), $V \subseteq U$ and $V \subseteq U'$, so $U = U'$, which is a contradiction. Similarly, in case (b), $V' \subseteq U$ and $V' \subseteq U'$, so $U = U'$, which is a contradiction.)

In fact, assuming this graph-theoretic condition, we can establish the following strengthening of Proposition 3.3. The proof also establishes several alternate characterizations of the condition that may be of independent interest.

**Proposition 3.4** Given random variables $X$ and $Y$, if $wpg(X, Y)$ does not contain a path of length 3, then $H(X \wedge Y) = M(X, Y)$.

The following figure shows that the converse of Proposition 3.4 is false, so the graph-theoretic restriction does not characterize when the equality holds.

| $S$ | $X_1$ | $X_2$ |
|---|---|---|
| 1 | C | G |
| 2 | C | U |
| 3 | A | G |
| 4 | G | C |
| 5 | U | A |
| 6 | A | U |

**Figure 3 – Path in Sequence Alignment**

Since $\Gamma_{X_1} = \{\{1,2\},\{3,6\},\{4\},\{5\}\}$ and $\Gamma_{X_2} = \{\{1,3\},\{2,6\},\{4\},\{5\}\}$, it follows from Proposition 2.3 that $\Gamma_{X_1} \vee \Gamma_{X_2} = \{\{1,2,3,6\},\{4\},\{5\}\}$. Hence $M(X_1, X_2) = H(X_1 \wedge X_2) = \log_2(6) - 4/3$, but

$$X_1^{-1}(\{\mathrm{C}\}) — X_2^{-1}(\{\mathrm{U}\}) — X_1^{-1}(\{\mathrm{A}\}) — X_2^{-1}(\{\mathrm{G}\}) — X_1^{-1}(\{\mathrm{C}\})$$

is a path of length 4 in $wpg(X_1, X_2)$.

## Section 4 – Discussion

As mentioned in the introduction, we have not seen an explicit reference to the Shannon order $\leq$ following the original definition in [10]. However, both the idea of zero conditional entropy used in Definition 1.1 ($H(X \mid Y) = 0$) and the equivalent notion of mutual information and entropy coinciding used in Proposition 3.1 ($M(X, Y) = H(X)$) have been studied in different contexts.

For example, universally quantified statements about zero conditional entropy play a prominent role in classical information theory ([1], pp. 49-52). By definition, a channel with output $Y$ is *deterministic* if $H(Y \mid X) = 0$ for every input $X$. In our setting, this is equivalent to saying that $Y \leq X$. Similarly, a channel is *lossless* if $H(X \mid Y) = 0$ for every input $X$, or equivalently, that $X \leq Y$. Therefore, a channel is *noiseless* (lossless and deterministic) if $X \equiv Y$ for every input $X$.

In addition, other papers have recognized the importance of the equality $M(X, Y) = H(X)$. For example, in [5], the idea is used as the basis of a methodology for predicting the regulatory structure of gene networks. Also, ([13], 8.1) essentially establishes Proposition 3.1(a) although he phrasing "$X$ is a *consequence* of $Y$" is used without explicitly defining the order. Other papers discuss the entropy of partitions and establish results that are related to ours. For example, ([12], Theorem 2.8) establishes a restricted version (using the assumptions of Section 2) of the fact noted at the end of Section 1 that $H(\bullet \mid X)$ is an order-preserving mapping. Also, based on Definition 1.1, ([12], Theorem 2.3) establishes the equivalence of (a) and (c) in Proposition 2.1.

The ideas presented in this paper also have some less evident connections with the comparative analysis of RNA sequences. Briefly, this technique uses the amount of covariation between positions in alignments of RNA sequences to predict base-pairs in the secondary structure ([8], chapter 7 in [7]). The idea of using mutual information to measure the covariation was introduced in [3]. This approach is related to the definition of *channel capacity* used in information theory.

Proposition 3.1 provides the connection with the present work. This result states that the mutual information between two positions is maximized precisely when the corresponding random variables are related in the order. Further details about these connections based on numerical experiments will be presented in [4].

It is well known that mutual information is a special case of relative entropy (based on (5) in Appendix 1). However, we do not seen any references that relate $M(X, Y)$ to either the behavior of a random variable (on the same sample space such as $X \wedge Y$) or to the properties of a graph (such as $wpg(X, Y)$). Either approach is potentially useful as a graph-theoretic alternative for studying the joint behavior of a pair of random variables. In particular, the proof of Proposition 3.4 is based on a lemma in Appendix 2 that characterizes when $wpg(X, Y)$ does not contain a path of length 3. It seems likely that the equivalence of parts (a) and (b) in the Lemma is well known, but we do not have a reference for it. Seemingly, this result must be generalized to characterize when $M(X, Y) = H(X \wedge Y)$ holds.

## References


1. Ash, R.B. Information Theory, *Dover Publications, Inc.*, New York, 1990.

2. Birkhoff, G.D. Lattice Theory, 3rd edition, *American Mathematical Society*, Providence, 1979.

3. Chiu, D.K.Y.; Kolodziejczak, T. Inferring consensus structure from nucleic acid sequences, *CABIOS* 7, 3, 1991, 347-352.

4. Che, Z.; Rice, M.D. An Order-theoretic Point of View on Comparative Sequence Analysis, manuscript in work.

5. D'haeseleer, P.; Liang S.; Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering, *Bioinformatics* 16 (8): 2000, 707-726.

6. Davey, B.A.; Priestley, H.A. Introduction to Lattices and Order, 2nd edition, *Cambridge University Press*, 2002.

7. Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

8. Gutell, R.R. Structure Prediction with Comparative Sequence Analysis – Rationale and Methods, manuscript posted at http://www.rna.icmb.utexas.edu/, 2000.

9. Lerman, I.C. Classification et analyse ordinale des données, *Dunod*, Paris, 1981.

10. Shannon, C.E. Collected Papers, Part A, article 62, Sloane, N.A.J. and Wyner, A.D. (editors), *IEEE Press*, 1993.

11. Sprinzl, M.; Horn, C.; Brown, M.; Ioudovitch, A.; Steinberg, S. Compilation of *tRNA* sequences and sequences of tRNA genes, Laboratory for Biochemistry, University of Bayreuth, Germany, http://www.uni bayreuth.de/departments/biochemie/trna/.

12. Simovici, D.; Jaroszewicz, S. Generalized Entropy and Decision Trees, *EGC 2003 - Journees francophones d'Extraction et de Gestion de Connaissances*, Lyon, France, January 2003, 369-380.

13. Topsøe, F. Basic Concepts, Identities and Inequalities - the Toolkit of Information Theory, *Entropy*, Issue 3, 2001, 162-190.

## Appendix 1 - Notation

This appendix provides the background for reading the technical aspects of the paper. The majority of the notation and results are found in chapter 1 of [1].

We assume that each random variable $X$ is defined on a sample space $S$ and has a finite set of values $\{x_1,\ldots, x_n\}$ that may be either numerical or symbolic. We assume that each sample space $S$ has a probability function $P$ defined on a family of subsets of $S$ that is closed under finite unions, finite intersections, and contains each set $X^{-1}(F)$ where $X$ is a random variable and $F \subseteq \{x_1,\ldots, x_n\}$. The probability that value $x_i$ occurs is denoted by

$$p(x_i) = P(X = x_i) = P(\{s \in S \mid X(s) = x_i\}).$$

The set $\{p(x_i) \mid 1 \le i \le n\}$ defines the *probability distribution* of $X$. We assume that $p(x_i) > 0$ for each value $x_i$.

The *entropy* of a random variable $X$ ([1], 1.2) is defined by the expression

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i)). \qquad \textbf{(1)}$$

In general, $H(X) \ge 0$ and if $X$ has $n$ possible values, then $H(X) \le log_2(n)$ ([1], 1.4.2). The value $H(X)$ can be interpreted as the average *uncertainty* associated with the set of events $\{X^{-1}(x_i) \mid 1 \le i \le n\}$. This uncertainty is removed by knowing the value of $X$.

Given random variables $X : S \to \{x_1,\ldots, x_n\}$ and $Y : S \to \{y_1,\ldots, y_m\}$, the random variable

$$(X, Y) : S \to \{(x_i, y_j) \mid 1 \le i \le n, 1 \le j \le m\}$$

is defined by $(X, Y)(s) = (X(s), Y(s))$ and has the probability distribution

$$p(x_i, y_j) = P(X = x_i \wedge Y = y_j) = P(\{s \in S \mid X(s) = x_i \wedge Y(s) = y_j\}).$$

The random variables $X$ and $Y$ are *independent* if $p(x_i, y_j) = p(x_i)p(y_j)$ for every pair of values $(x_i, y_j)$.

The *joint entropy* of the random variable $(X, Y)$ ([1], 1.4) is defined by the expression

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2(p(x_i, y_j)). \qquad \textbf{(2)}$$

In general, $H(X, Y) = H(Y, X)$ and $H(X, Y) \le H(X) + H(Y)$ with equality holding if $X$ and $Y$ are independent ([1], 1.4.3).

Given random variables $X : S \to \{x_1,\ldots, x_n\}$ and $Y : S \to \{y_1,\ldots, y_m\}$, the *conditional probability* of value $x_i$ given value $y_j$ is defined by

$$p(x_i \mid y_j) = p(x_i, y_j)/p(y_j).$$

The *conditional entropy* of $X$ given $Y$ ([1], 1.4) is defined by the expression

$$H(X \mid Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2(p(x_i \mid y_j)). \qquad \textbf{(3)}$$

In general, $H(X \mid Y) \geq 0$ and $H(X \mid Y) \leq H(X)$ with equality holding if and only if $X$ and $Y$ are independent ([1], 1.4.5). This inequality says that knowing the value of $Y$ may decrease the uncertainty about $X$. One can also establish the following equality ([1], 1.4.4):

$$H(X, Y) = H(Y) + H(X \mid Y). \qquad \textbf{(4)}$$

Given random variables $X$ and $Y$, the *mutual information* between $X$ and $Y$ is defined by the expression

$$M(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2\left[\frac{p(x_i, y_j)}{p(x_i)\,p(y_j)}\right]. \qquad \textbf{(5)}$$

In general, $M(X, Y) \geq 0$ and $M(X, Y) = M(Y, X)$ ([1], 1.5). The mutual information can be interpreted as a measure of the *covariation* of $X$ and $Y$, that is, the degree to which changes in the values of one random variable are reflected by changes in the other. In particular, if $X$ and $Y$ are independent, then $M(X, Y) = 0$.

Using (4) and (5), one can establish that

$$M(X, Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X \mid Y). \qquad \textbf{(6)}$$

[Actually, (6) is the definition of mutual information used in ([1], 1.5.1); a subsequent discussion establishes the equivalence of (5) and (6).]

## Appendix 2 – Proofs

This appendix provides the detailed proofs of the results found in the paper.

PROOF 1.1

(a) → (b): By Definition 1.1,

$$(\#) \qquad H(X \mid Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2(p(x_i \mid y_j)) = 0.$$

By assumption, for each value $y_j$ of $Y$,

$$(\#\#) \qquad p(y_j) = \sum_{i=1}^{n} p(x_i, y_j) > 0$$

so there exists a value $x_i$ of $X$ that satisfies $p(x_i, y_j) > 0$. Hence $p(x_i \mid y_j) = p(x_i, y_j) / p(y_j) > 0$, so it follows from (#) that $p(x_i \mid y_j) = 1$. Therefore, it follows from (##) that $x_i$ must be unique.

(b) → (c): Define the mapping $\varphi : \{y_1,\ldots, y_m\} \to \{x_1,\ldots, x_n\}$ by $\varphi(y_j) = x_i$, where $p(x_i \mid y_j) = 1$. It follows that $p(x_i, y_j) = p(x_i | y_j)\, p(y_j) = p(y_j)$; hence by definition,

$$P(X = \varphi(y_j) \wedge Y = y_j) = P(X = x_i \wedge Y = y_j) = p(x_i, y_j) = p(y_j) = P(Y = y_j).$$

To show that φ is a surjection, suppose $x_i$ is a value of $X$. Since

$$p(x_i) = \sum_{j=1}^{m} p(x_i, y_j) > 0$$

there exists a value $y_j$ of $Y$ such that $p(x_i, y_j) > 0$. Since $p(\varphi(y_j) \mid y_j) = 1$, it follows from (##) that $p(x_k \mid y_j) = 0$ if $\varphi(y_j) \neq x_k$; hence $\varphi(y_j) = x_i$, so φ is a surjection.

(c) → (d): For each $1 \le j \le m$, define $S_j = Y^{-1}(y_j) \cap X^{-1}(\varphi(y_j))$ and let $S' = \bigcup_{j=1}^{m} S_j$ .

By (c), $P(S_j) = P(Y = y_j)$ for each $1 \le j \le m$. Since $\{S_j\}$ is a disjoint family of sets,

$$P(S') = \sum_{j=1}^{m} P(S_j) = \sum_{j=1}^{m} p(y_j) = 1.$$

By definition, the equality $X = \varphi \circ Y$ holds for every member of $S'$; hence $P(X = \varphi \circ Y) \ge P(S') = 1$, so (d) holds.

(d) → (a): Suppose $p(x_i, y_j) > 0$ for values $x_i$ of $X$ and $y_j$ of $Y$. By (d),

$$p(x_i, y_j) = P(X = x_i \wedge Y = y_j) = P(\varphi \circ Y = x_i \wedge Y = y_j).$$

Since $p(x_i, y_j) > 0$, $\varphi(y_j) = x_i$; hence $P(\varphi \circ Y = x_i \wedge Y = y_j) = P(Y = y_j) = p(y_j)$. Therefore, by the above equality, $p(x_i \mid y_j) = p(x_i, y_j)/p(y_j) = 1$. It follows from (#) that $H(X \mid Y) = 0$; hence $X \le Y$.

□

PROOF 1.2

By Proposition 1.1(b), $X \equiv X$ for each random variable X; hence $\equiv$ is a reflexive relation. By definition, $\equiv$ is also symmetric. The fact that $\equiv$ is transitive is a consequence of the following result:

(&) Given random variables $X$, $Y$, and $Z$, if $X \le Y$ and $Y \le Z$, then $X \le Z$.

Assume that $X \le Y$ and $Y \le Z$. By Proposition 1.1(d), there exist surjections $\varphi : \{y_j\} \to \{x_i\}$ and $\psi : \{z_k\} \to \{y_j\}$ such that

$$(\#) \quad P(X = \varphi \circ Y) = P(Y = \psi \circ Z) = 1.$$

Define $\tau = \varphi \circ \psi$ and the sets $S_1 = \{s \in S \mid X(s) = \varphi(Y(s))\}$ and $S_2 = \{s \in S \mid Y(s) = \psi(Z(s))\}$.

By (#), $P(S_1) = P(S_2) = 1$, so $P(S_1 \cap S_2) = 1$ and for each $s \in S_1 \cap S_2$, $X(s) = \varphi(Y(s)) = \varphi(\psi(Z(s))) = \tau(Z(s))$. Therefore, $P(X = \tau \circ Z) = 1$, so by Proposition 1.1(d), $X \le Z$.

□

PROOF 1.3

It follows from assertion (&) in the proof of 1.3 that $\le$ is a well-defined transitive relation on $L$. Clearly, it is also reflexive since $X \le X$ holds for each random variable $X$ and it is anti-symmetric based on the definition of $\equiv$. Therefore, $\le$ is a partial order on $L$.

Suppose $X$ and $Y$ are random variables and $X \le Y$. By definition, $H(X \mid Y) = 0$, so by equation (4) in Section 0, $H(X, Y) = H(Y) + H(X \mid Y) = H(Y)$. Similarly, $H(X, Y) = H(X) + H(Y \mid X) \ge H(X)$ since $H(Y \mid X) \ge 0$, so it follows that $H(X) \le H(X, Y) = H(Y)$.

Since $H$ preserves the order $\le$ on random variables, it is a well-defined order-preserving mapping on $L$.

□

PROOF 1.4

Let $F = \{x_i\}$ and $G = \{y_j\}$ denote the values of $X$ and $Y$, respectively. Then the product set $F \times G$ denotes the values of $Z = (X, Y)$. It is straightforward to verify that the projection mappings $F \times G \to F$ and $F \times G \to G$ satisfy condition (d) in Proposition 1.1, so $X \le Z$ and $Y \le Z$.

Let $U$ be a random variable with values $H = \{u_k\}$ that satisfies $X \le U$ and $Y \le U$. By Proposition 1.1(d), there exist surjections $\varphi : H \to F$ and $\psi : H \to G$ satisfying

$$(\#) \quad P(X = \varphi \circ U) = P(Y = \psi \circ U) = 1.$$

Define the surjection $\tau : H \to F \times G$ by $\tau(u_k) = (\varphi(u_k), \psi(u_k))$ and define $S_1 = \{s \in S \mid X(s) = \varphi(U(s))\}$ and $S_2 = \{s \in S \mid Y(s) = \psi(U(s))\}$.

By (#), $P(S_1) = P(S_2) = 1$, so $P(S_1 \cap S_2) = 1$. For each $s \in S_1 \cap S_2$, $X(s) = \varphi(U(s))$ and $Y(s) = \psi(U(s))$, so $Z(s) = \tau(U(s))$. Therefore, $P(Z = \tau \circ U) = 1$, so by Proposition 1.1(d), $Z \le U$. Hence $Z$ is the least upper bound of $X$ and $Y$ in $L$.

□

PROOF 1.5

Assume that $X \le Y$ and let $F = \{x_i\}$ and $G = \{y_j\}$ denote the values of $X$ and $Y$, respectively. By Proposition 1.1(d), there exists a surjection $\varphi : G \to F$ such that $P(X = \varphi \circ Y) = 1$. Since $\varphi$ is a surjection, there exists a mapping $\alpha : F \to G$ such that $\varphi \circ \alpha = id_F$ (the identity mapping on $F$). Let $\lambda = \alpha \circ \varphi$ and define $X^* = \lambda \circ Y$. We claim that $X \equiv X^*$.

Since

$$P(X^* = \alpha \circ X) = P(\{s \in S \mid (\alpha \circ \varphi \circ Y)(s) = (\alpha \circ X)(s)\})$$
$$\geq P(\{s \in S \mid (\varphi \circ Y)(s) = X(s)\}) = 1,$$

$P(X^* = \alpha \circ X) = 1$. By definition, $\alpha$ is a surjection onto the values of $X^*$, so it follows that $X^* \leq X$.

Also,

$$P(X = \varphi \circ X^*) = P(\{s \in S \mid X(s) = (\varphi \circ \alpha \circ \varphi \circ Y)(s)\})$$
$$= P(\{s \in S \mid X(s) = (\varphi \circ Y)(s)\}) = 1$$

since $\varphi \circ \alpha = id_F$. Because $\alpha$ is a bijection (onto the values of $X^*$) and $\varphi \circ \alpha = id_F$, $\varphi$ restricted to the values of $X^*$ is a surjection onto $F$, so it follows that $X \leq X^*$. Hence $X \equiv \lambda \circ Y$.

Conversely, suppose $X \equiv \lambda \circ Y$. Then by definition, $X \leq \lambda \circ Y$, so by Proposition 1.1(d), there exists a mapping $\varphi$ such that $P(X = \alpha \circ Y) = 1$ where $\alpha = \varphi \circ \lambda$.

Suppose $p(x_i, y_j) > 0$ for values $x_i$ of $X$ and $y_j$ of $Y$. Then

$$p(x_i, y_j) = P(X = x_i \wedge Y = y_j) = P(\alpha \circ Y = x_i \wedge Y = y_j).$$

Since $p(x_i, y_j) > 0$, $\alpha(y_j) = x_i$; hence $P(\alpha \circ Y = x_i \wedge Y = y_j) = P(Y = y_j) = p(y_j)$. Therefore, by the above equality, $p(x_i \mid y_j) = p(x_i, y_j)/p(y_j) = 1$. Hence it follows from Proposition 1.1(b) that $X \leq Y$.

□

PROOF 1.6

The set **F** is non-empty since it contains a member $A_r$ that satisfies $[A_r] = \perp$. Also, by Proposition 1.6, **F** is finite. Therefore, by Proposition 1.5, the element $Z = \vee\mathbf{F}$ exists.

Since each $A_r$ satisfies $A_r \leq X$ and $A_r \leq Y$, $Z \leq X$ and $Z \leq Y$. Suppose $U \leq X$ and $U \leq Y$. By definition, there exists an $A_r$ such that $A_r \equiv U$. Hence $U \leq A_r$ and $A_r \leq Z$ implies that $U \leq Z$. Therefore, $Z$ is the largest random variable satisfying $Z \leq X$ and $Z \leq Y$.

□

PROOF 2.1

(a) $\leftrightarrow$ (b): Since $P(\{s\}) > 0$ for each event $s$, $S$ is the only set that satisfies $P(S) = 1$. Hence $P(X = \varphi \circ Y) = 1$ is equivalent to saying that $X = \varphi \circ Y$. Therefore, by Proposition 1.1, (a) and (b) are equivalent.

(b) $\to$ (c): Choose $Y^{-1}(y_j) \in \Gamma_Y$ and let $x_i = \varphi(y_j)$. By (b), $Y^{-1}(y_j) \subseteq Y^{-1}(\varphi^{-1}(x_i)) = X^{-1}(x_i)$, where $X^{-1}(x_i) \in \Gamma_X$. Hence $\Gamma_Y < \Gamma_X$.

(c) → (b): Given $Y^{-1}(y_j) \in \Gamma_Y$, choose $x_i = \varphi(y_j)$ such that $Y^{-1}(y_j) \subseteq X^{-1}(x_i)$. This association defines a surjection $\varphi : \{y_1,\ldots, y_m\} \to \{x_1,\ldots, x_n\}$. (Given a value $x_i$ of $X$, choose $s \in X^{-1}(x_i)$ and select a set $Y^{-1}(y_j) \in \Gamma_Y$ that contains $s$. Since $\Gamma_X$ is a partition, $Y^{-1}(y_j) \subseteq X^{-1}(x_i)$, so $\varphi(y_j) = x_i$.) Based on the definition of φ, it follows that $X = \varphi \circ Y$.

□

PROOF 2.2

Define the mapping $g : L \to L(S)$ by $g([X]) = \Gamma_X$. This mapping is well-defined since if $X \equiv Y$, then by Proposition 2.1, $\Gamma_X < \Gamma_Y$ and $\Gamma_Y < \Gamma_X$, so $\Gamma_X = \Gamma_Y$.

Also, $g$ is one-to-one since if $g([X]) = \Gamma_X = g([Y]) = \Gamma_Y$, then by Proposition 2.1, $X \leq Y$ and $Y \leq X$. Hence $X \equiv Y$, so $[X] = [Y]$.

To show that $g$ is a surjection, assume that $\{A_i\}$ is a partition of $S$. Define the random variable $X : S \to \{x_i\}$ by $X(s) = x_i$ if $s \in A_i$; then $\Gamma_X = \{A_i\}$.

Since $g$ is a bijection, Proposition 2.1 establishes that $g$ is an order-isomorphism, so it is also a lattice isomorphism ([6], 2.19(ii)).

□

PROOF 2.3

By definition, Γ is a partition of $S$ that satisfies $\Gamma_X < \Gamma$ and $\Gamma_Y < \Gamma$. Suppose $\Gamma_X = \{A_i\}$, $\Gamma_Y = \{B_j\}$, and Φ is a partition of $S$ that satisfies $\Gamma_X < \Phi$ and $\Gamma_Y < \Phi$. Let $C$ be a connected component of $wpg(X, Y)$. If $A_i \subseteq V$ and $B_j \subseteq W$ for $V, W \in \Phi$ and $A_i \cap B_j \neq \varnothing$, then $V = W$. Hence if $(A_i, B_j)$ is an edge, there exists $V \in \Phi$ such that $A_i \cup B_j \subseteq V$. Since any two vertices in $C$ are connected by a sequence of edges, it follows that $\cup C \subseteq V$ for some $V \in \Phi$. Hence $\Gamma < \Phi$ which establishes that $\Gamma = \Gamma_X \vee \Gamma_Y$.

□

PROOF 3.2

Suppose $A \leq B$. By (6) in Appendix 1,

$$\begin{aligned} M(B, X) - M(A, X) &= H(X) - H(X \mid B) - (H(X) - H(X \mid A)) \\ &= H(X \mid A) - H(X \mid B). \end{aligned}$$

Therefore, it suffices to show that $H(X \mid B) \leq H(X \mid A)$. Based on Proposition 2.1, this result is stated in ([1], exercise 3.4) and ([12], Theorem 2.5). Here we present a proof based on Proposition 2.1. By that result, there exists a surjection $\varphi : \{b_1,\ldots, b_m\} \to \{a_1,\ldots, a_n\}$ such that $A = \varphi \circ B$. Then by definition

$$\begin{aligned} H(X \mid A) &= -\sum\{p(x_i,a_j) \log_2(p(x_i \mid a_j)) \mid i, j\} \\ &= -\sum\{p(x_i,a_j) \log_2(p(x_i \mid \varphi(b_k))) \mid i, j, k\} \\ &= -\sum\{p(x_i,a_j,b_k) \log_2(p(x_i \mid \varphi(b_k))) \mid i, j, k\} \end{aligned}$$

since $\sum\{p(x_i,a_j,b_k) \mid k\} = p(x_i,a_j)$ and $\sum\{p(x_i,a_j,b_k) \mid j\} = p(x_i,b_k)$.

Therefore,

$$\begin{aligned} H(X \mid A) &= -\sum\{p(x_i,b_k) \log_2(p(x_i \mid \varphi(b_k))) \mid i, k\} \\ &= -\sum\{p(b_k)\, p(x_i \mid b_k) \log_2(p(x_i \mid \varphi(b_k))) \mid i, k\} \\ &= \sum\{p(b_k)\, [-\sum \{p(x_i \mid b_k) \log_2(p(x_i \mid \varphi(b_k))) \mid i\}] \mid k\} \\ &\geq \sum\{p(b_k)\, [-\sum \{p(x_i \mid b_k) \log_2(p(x_i \mid b_k)) \mid i\}] \mid k\} \end{aligned}$$

since

$$-\sum\{p(x_i \mid b_k)\log_2(p(x_i \mid \varphi(b_k))) \mid i\} \geq -\sum\{p(x_i \mid b_k)\log_2(p(x_i \mid b_k)) \mid i\}$$

by Gibbs Lemma ([1], Lemma 1.4.1).

Hence $H(X \mid A) \geq -\sum\{p(x_i,b_k)\ln(p(x_i \mid b_k)) \mid i, k\} = H(X \mid B)$.

□

PROOF 3.4

We will use the following result (that is established below):

LEMMA The following statements are equivalent:

(a) The graph $wpg(X, Y)$ does not contain a path of length 3.

(b) For each $U \in \Gamma_X$ and $V \in \Gamma_Y$, either $U \cap V = \emptyset$, $U \subseteq V$, or $V \subseteq U$.

(c) The covers $\Gamma_X$ and $\Gamma_Y$ can be written in the following form:

$$\Gamma_X = \{A_i \mid 1 \leq i \leq p\} \cup \{A_F^*\} \cup \{C_k\}$$
$$\Gamma_Y = \{B_j \mid 1 \leq j \leq q\} \cup \{B_G^*\} \cup \{C_k\}$$

where $F \subseteq \{1, ..., q\}$, $G \subseteq \{1, ..., p\}$, $A_F^* = \cup\{B_j \mid j \in F\}$, $B_G^* = \cup\{A_i \mid i \in G\}$, $A_i \cap B_j = \emptyset$, $\cup A_i = \cup B_G^*$, and $\cup B_j = \cup A_F^*$.

If $wpg(X, Y)$ does not contain a path of length 3, then based on the descriptions given in part (c), it is straightforward to verify that $\Gamma_X \wedge \Gamma_Y = \{A_i\} \cup \{B_j\} \cup \{C_k\}$ and $\Gamma_X \vee \Gamma_Y = \{A_F^*\} \cup \{B_G^*\} \cup \{C_k\}$.

Let $N = |S|$. Then

$$H(X) = -\left[\sum |A_i|/N \log_2(|A_i|/N) + \sum |A_F^*|/N \log_2(|A_F^*|/N) + \sum |C_k|/N \log_2(|C_k|/N)\right]$$
$$H(Y) = -\left[\sum |B_j|/N \log_2(|B_j|/N) + \sum |B_G^*|/N \log_2(|B_G^*|/N) + \sum |C_k|/N \log_2(|C_k|/N)\right].$$

Also, using the description of $\Gamma_X \wedge \Gamma_Y$ based on Propositions 1.4 and 2.2,

$$H(X \vee Y) = -\left[\sum |A_i|/N \log_2(|A_i|/N) + \sum |B_j|/N \log_2(|B_j|/N) + \sum |C_k|/N \log_2(|C_k|/N)\right].$$

Therefore, by (6) in Appendix 1,

$$\begin{aligned} M(X, Y) &= H(X) + H(Y) - H(X, Y) \\ &= -\left[\sum |A_F^*|/N \log_2(|A_F^*|/N) + \sum |B_G^*|/N \log_2(|B_G^*|/N) + \sum |C_k|/N \log_2(|C_k|/N)\right]. \end{aligned}$$

Based on the description of $\Gamma_X \vee \Gamma_Y$ given in Proposition 2.3, the preceding expression equals $H(X \wedge Y)$ which completes the proof.

PROOF LEMMA

(a) $\leftrightarrow$ (b): If (b) does not hold, there exist $U_1 \in \Gamma_X$ and $V_1 \in \Gamma_Y$ such that $U_1 \cap V_1 \neq \emptyset$, $U_1 \setminus V_1 \neq \emptyset$, and $V_1 \setminus U_1 \neq \emptyset$. Since $\Gamma_X$ and $\Gamma_Y$ are covers, there exist $U_2 \in \Gamma_X$ and $V_2 \in \Gamma_Y$ such that $U_1 \neq U_2$, $U_2 \cap V_1 \neq \emptyset$, $V_1 \neq V_2$, and $U_1 \cap V_2 \neq \emptyset$. Then the sets $V_2$ — $U_1$ — $V_1$ — $U_2$ define a path of length 3 in $wpg(X, Y)$.

If (a) does not hold, suppose the sets $V_2$ — $U_1$ — $V_1$ — $U_2$ define a path of length 3 in *wpg*($X$, $Y$). Then $U_1 \cap V_1 \neq \varnothing$, $U_1 \cap V_2 \neq \varnothing$, and $U_2 \cap V_1 \neq \varnothing$. Hence $V_1$ is not a subset of $U_1$ and $U_1$ is not a subset of $V_1$, so condition (b) does not hold for the pair $U_1$, $V_1$.

(b) $\leftrightarrow$ (c): Based on the description in (c), one can verify (c) $\rightarrow$ (b). To establish the converse, define the families

$$\Phi = \{U \in \Gamma_X \mid U \subset V \text{ for some } V \in \Gamma_Y\}$$
$$\Omega = \{V \in \Gamma_Y \mid V \subset U \text{ for some } U \in \Gamma_X\}.$$

<u>Claim 1</u> (i) $U \in \Gamma_X \backslash \Phi \;\Rightarrow\; U \in \Gamma_Y$ or $U = \cup\{V \in \Gamma_Y \mid V \subset U\}$.

(ii) $V \in \Gamma_Y \backslash \Omega \;\Rightarrow\; V \in \Gamma_X$ or $V = \cup\{U \in \Gamma_X \mid U \subset V\}$.

<u>PROOF</u>

(i) Choose $U \in \Gamma_X \backslash \Phi$ and define $F = \{V \in \Gamma_Y \mid U \cap V \neq \varnothing\}$. It follows from (b) that for each $V \in F$, either $U \subseteq V$ or $V \subseteq U$. Since $U \notin \Phi$, if $U \notin \Gamma_Y$, then it follows that $V \subset U$ for each $V \in F$. Since $\Gamma_Y$ is a cover, it follows that $U = \cup\{V \in \Gamma_Y \mid V \subset U\}$. Claim (ii) is established in a similar manner.

Let $\Phi = \{A_i \mid 1 \leq i \leq p\}$, $\Omega = \{B_j \mid 1 \leq j \leq q\}$, and $\Gamma_X \cap \Gamma_Y = \{C_k\}$.

<u>Claim 2</u> The families $\{A_i\}$, $\{B_j\}$, and $\{C_k\}$ are disjoint collections of sets.

<u>PROOF</u> Based on (b), if $A_i \cap C_k \neq \varnothing$, then either $A_i \subseteq C_k$ or $C_k \subseteq A_i$. By definition, there exists $V \in \Gamma_Y$ such that $A_i \subset V$; hence $C_k \subseteq A_i$ cannot hold. If $A_i \subseteq C_k$, then $C_k \in \Gamma_X$ implies $A_i = C_k$, which also cannot hold. Hence $A_i \cap C_k = \varnothing$. A similar argument shows that $B_j \cap C_k = \varnothing$.

If $A_i \cap B_j \neq \varnothing$, choose $U \in \Gamma_X$ such that $B_j \subset U$. Since $\Gamma_X$ is a partition, it follows that $A_i = U$, so $B_j \subset A_i$. A similar argument shows that $A_i \subset B_j$, which gives a contradiction.

Based on Claim 1, if $U \in \Gamma_X \backslash \{A_i\}$, then either $U = C_k$ for some $k$ or $U = \cup\{V \in \Gamma_Y \mid V \subset U\}$. Since by definition, $\{V \in \Gamma_Y \mid V \subset U\} \subseteq \{B_j\}$, it follows that $U = A_F{}^*$ for some $F \subseteq \{1, ..., q\}$. Therefore, $\Gamma_X$ has the form specified in (c). It also follows from Claim 2 that $A_i \cap B_j = \varnothing$, $\cup A_i = \cup B_G{}^*$, and $\cup B_j = \cup A_F{}^*$. A similar argument shows that $\Gamma_Y$ has the form specified in (c).

□