entropy

MDPI

Article

Information Theory in Computational Biology: Where We Stand Today

Pritam Chanda [1,2,*], Eduardo Costa [3], Jie Hu [1], Shravan Sukumar [1], John Van Hemert [4,*] and Rasna Walia [4]

1 Corteva Agriscience™, Indianapolis, IN 46268, USA; jie.hu@corteva.com (J.H.); shravan.sukumar@corteva.com (S.S.)
2 Computer and Information Science, Indiana University-Purdue University, Indianapolis, IN 46202, USA
3 Corteva Agriscience™, Mogi Mirim, Sao Paulo 13801-540, Brazil; eduardo.costa@corteva.com
4 Corteva Agriscience™, Johnston, IA 50131, USA; rasna.walia@corteva.com
* Correspondence: pritam.chanda@corteva.com (P.C.); john.vanhemert@corteva.com (J.V.H.)

Received: 30 April 2020; Accepted: 3 June 2020; Published: 6 June 2020

Abstract: "A Mathematical Theory of Communication" was published in 1948 by Claude Shannon to address the problems in the field of data compression and communication over (noisy) communication channels. Since then, the concepts and ideas developed in Shannon's work have formed the basis of information theory, a cornerstone of statistical learning and inference, and has been playing a key role in disciplines such as physics and thermodynamics, probability and statistics, computational sciences and biological sciences. In this article we review the basic information theory based concepts and describe their key applications in multiple major areas of research in computational biology—gene expression and transcriptomics, alignment-free sequence comparison, sequencing and error correction, genome-wide disease-gene association mapping, metabolic networks and metabolomics, and protein sequence, structure and interaction analysis.

Keywords: information theory; entropy; computational biology; gene expression; transcriptomics; sequence comparison; error correction; disease-gene association mapping; metabolic networks; metabolomics; protein structure; interaction analysis

1. Introduction

Information theory has its roots in communication systems that deal with data compression and coding theorems for transmission of information from one source to another over noisy channels. Claude Shannon's seminal work in communication theory from 1948 [1] provided the mathematical foundations for quantification and representation of information that made today's digital era possible. It introduced the concept of channel capacity, defining the amount of information that can be sent over a noisy channel, bounded by the maximum possible transmission rate—"Shannon's Limit" and stated that it is possible to transmit information through a noisy channel at a rate less than the maximum channel capacity keeping the probability of error at the receiver's end arbitrarily small [2]. Since its inception, the significance and potentials of Shannon's results were rapidly recognized, leading to further developments in probabilistic information measures and its wide application in diverse theoretical and applied fields outside of its original scope and intent in communication theory. Indeed, quantification and reduction of noise and uncertainty from data observations is key to improving statistical inference and it is not just limited to communication theory but any area that deals with noisy data. Biology is perhaps one of the experimental areas that is permeated by noise and variability in all observational levels, ranging from the most basic molecular and subcellular processes such as gene expressions and signaling pathways to complex interactions and dynamics of tissues, organs, organisms and populations [3]. Therefore, it is not surprising that information theory has found

many theoretical advancements and witnessed myriad applications dealing with biological data. This is particularly true in the umbrella field of computational biology and bioinformatics that deals with computational applications of mathematical and statistical methods in the study of biological systems and processes. In this domain, information theory is widely used for model development and data analysis for a variety of biologically derived data types ranging from molecular, sequence and phenotypic data in genomics and genetics to gene expression, protein and spectral data in transcriptomics, proteomics and metabolomics, respectively [4–11].

Despite the numerous applications of information theoretic methods in many areas within computational biology, there have been only a few recent articles comprehensively reviewing theoretical developments and methodological applications to address problems in biology. Applications of information theory at the molecular level were reviewed in 2010 in [12] where the authors investigated information theory in the context of quantifying information in DNA-binding sites and the study of protein-DNA interactions and showed how coding theory can be used to explain molecular efficiency. In 2014, Vinga et al. [6] reviewed the application of information theory focusing exclusively on biological sequences. Their article discussed both global sequence analysis methods using metrics such as block-entropy (or L-tuple entropy) [13–16] that uses chaos game representation of genomic sequences [17,18], and local sequence analysis techniques such as classification of motifs, prediction of transcription factor binding sites and sequence characterization inspired by information theoretic frameworks with natural language underpinnings. More recently in 2016, information theory based methods to understand and quantify signal transmission in cellular systems were reviewed in [5]. Entropy based analysis has been widely applied in characterizing cell signal transduction and associated biochemical pathways that are inherently noisy owing to many factors such as the stochastic fluctuations in genetic circuits and intrinsic promiscuity of protein–protein interactions [19]. A review of multiple information theory based methods to quantify information processing at the level of individual cells in the context of biochemical networks with non-linear dynamics can be found in [5,7]. The article [5] also provided an excellent summary of entropy based methods investigating protein-interaction networks, while an earlier review authored by the same group [4] discussed information theory usage in gene regulatory and metabolic networks.

In this review article, we aim for the following—(1) we revisit some topics from previous reviews covering key newer entropy and information theory based approaches in those areas (gene regulatory networks, protein–protein interaction and metabolic networks); (2) we discuss information theory based measures of multivariate gene–gene interactions and survey articles using them in genome-wide disease-gene association mapping; (3) we offer a broad summary of information theory based applications by including discussion on several key and recent uses of information theory in topics within computational biology that were not collectively reviewed before (such as protein structure and interaction analysis, protein coevolution, sequencing error correction, alignment-free phylogeny, optimization and dimensionality reduction in biology).

The remainder of the article is organized into two major parts. In the first part, we introduce the readers to basic concepts, quantities and notations in information theory that form the basis of many advanced metrics and information theoretic algorithms discussed later. In the second part, we introduce and elaborate upon key applications of information theory in specific areas within computational biology mentioned in the preceding paragraph, focusing on more recent developments. Any application specific metrics and extensions and generalizations of the basic concepts will be discussed within the context of each application. Because of its mathematical nature, the first part will help to provide a uniform vocabulary and mathematical symbols to explain the applications discussed in the second part of the article.

2. Basic Metrics in Information Theory

2.1. Self-Information and Entropy

Shannon's entropy [2] constitutes the basic building block for the information theoretic metrics to be discussed in the article. It is easiest to comprehend entropy when described in the context of

uncertainty in discrete random variables. The primary inspiration behind entropy is based on the idea of information content of a probabilistic event. An event that occurs with high probability has less information content than an event that occurs with lesser probability. So, learning the occurrence of a less likely event such as "solar eclipse today" is more informative than knowing about a more likely event "rainy today". This idea can be formally described using the metric of self-information (SI) that obeys the intuition of information content of an event. Given a discrete random variable X that assumes values from the set $V_X = \{x_1, x_2, \ldots, x_N\}$ and follows a probability distribution P_X, SI for a single event $X = x_i$ is defined as

$$SI_X(x_i) = -\log[P_X(x_i)] \tag{1}$$

Clearly, if an event $X = x_i$ occurs with certainty ($P_X(x_i) = 1$)), $SI_X(x_i) = 0$. This definition can be extended to an event pair involving two discrete random variables X and Y. For the joint occurrence of the events $X = x_i$ and $Y = y_j$ as:

$$SI_{X,Y}(x_i, y_j) = -\log[P_{X,Y}(x_i, y_j)] \tag{2}$$

where Y assumes values from the set $V_Y = \{y_1, y_2, \ldots, y_N\}$, follows a probability distribution P_Y, and $P_{X,Y}(x_i, y_j)$ is the joint probability of the two events $X = x_i$ and $Y = y_j$. This also highlights that when events described by $X = x_i$ and $Y = y_j$ are independent, the joint distribution should factor as $P_{X,Y}(x_i, y_j) = P_X(x_i)P_Y(y_j)$, so that the self-information $SI_{X,Y}(x_i, y_j) = SI_X(x_i) + SI_Y(y_j)$ is additive. Entropy (denoted mathematically by the symbol H) builds on the concept of self-information for a single event. For random variable X, it formalizes the information content of all the events $X = x_i$ ($i = 1, \ldots, N$) by taking the expectation with respect to the probability distribution of X:

$$H(X) = -\sum_{x_i \in V_X} P_X(x_i)\log[P_X(x_i)] = E_X(-\log[P_X]) \tag{3}$$

Entropy, thus, represents the uncertainty in X or information gained by observing a random variable X following the distribution P_X. Because it is an expectation, it depends on the distribution of the random variable X rather than an observed value. In other words, it represents the expected amount of information in an event occurrence following the distribution P_X and provides a lower bound to the number of bits required (when log is base 2) on average to encode each event drawn from the distribution. If some events are much more likely than others, entropy will be low. For distributions that are close to uniform, entropy is high as uncertainty is high. Analogous definition exists for a continuous random variable X, where the entropy (also referred to as cross-entropy or differential entropy) is defined as,

$$H(X) = -\int P_X(x_i)\log[P_X(x_i)]dx_i \tag{4}$$

where P refers to the continuous probability distribution of X.

As an example, consider X to be a random variable representing the alleles of a single nucleotide polymorphism (SNP) in the human genome. Assuming the SNP is bi-allelic, it can have 0, 1 or 2 copies of the minor allele in a genome. So X can take values from set $\{0, 1, 2\}$ representing the number of copies of the minor allele in a genome at that position. If we identify the minor allele counts for the SNP across N genomes, $P_X(x = i)$, $i \in \{0, 1, 2\}$ will give the respective probabilities for each of the three minor allele counts, and can be computed as $P_X(x = i) = n_i/N$, where n_i is the frequency of minor allele count i and N is the total number of genomes. Then the entropy can be computed as $H(X) = -\sum_{i\in\{0,1,2\}} P_X(x = i)\log[P_X(x = i)]$. It will be zero when the SNP has a single allele; it is maximized when $P_X(x)$ has a uniform distribution, i.e., $P_X(x = i) = 1/3$, $i \in \{0, 1, 2\}$.

The above definitions can be extended to multiple random variables. With a slight abuse of notation, for multiple discrete random variables X_1,...,X_M, entropy is defined using their joint probability distribution $P_{X_1,\ldots,X_M}$ as,

$$H(X_1,\ldots,X_M) = -\sum_{x_1\in V_{X_1}} \ldots \sum_{x_M\in V_{X_M}} P_{X_1,\ldots X_M}(x_1,\ldots,x_M)\log[P_{X_1,\ldots X_M}(x_1,\ldots,x_M)] \quad (5)$$

2.2. Conditional Entropy

Given two random variables X and Y, the entropy of X given the event $Y = y_i$ is defined $H(X|Y = y_i) = -\sum_{x_k\in V_X} P_{X|Y}(x_k|y_i)\log[P_{X|Y}(x_k|y_i)] = E_{X|Y}(-\log[P_{X|Y}])$ from which the conditional entropy of X given Y is defined as [2]

$$H(X|Y) = \sum_{y_i\in V_Y} P_Y(y_i)H(X|Y = y_i) = E_Y(E_{X|Y}(-\log[P_{X|Y}])) \quad (6)$$

Continuing with the example of the SNP from above, assume X is the SNP random variable and Y is a random variable associated with a measured phenotype, say presence ($Y = 1$) or absence ($Y = 0$) of a diseased condition. Assume n_{ij} be the frequency of minor allele count $i \in \{0, 1, 2\}$ for all genomes with a given disease status $Y = j$, $j \in \{0, 1\}$. Also assume c_j be the frequency of disease status $Y = j$ and N be the total number of genomes. Then the empirical conditional probability distribution is given by $P_{X|Y}(X = i|Y = j) = n_{ij}/c_j$; the conditional entropy can be calculated as,

$$\begin{aligned} H(X|Y) &= \sum_{j\in\{0,1\}} P_Y(Y = j)H(X|Y = j) \\ &= -\sum_{j\in\{0,1\}} P_Y(Y = j) \sum_{i\in\{0,1,2\}} P_{X|Y}(X = i|Y = j)\log[P_{X|Y}(X = i|Y = j)] \end{aligned}$$

The marginal distribution of Y is given by $P_Y(Y = j) = c_j/N$.

2.3. Relative Entropy

While Shannon's definition of entropy provides the average information content in a given probability distribution, it is often necessary to obtain information content of the occurrence of events in a probability distribution P_X with respect to a reference probability distribution Q_X, which is given by relative entropy or Kullback–Leibler Divergence [2,20]. It can be thought of as a measure of the distance between two distributions and measures the loss in information of using probability distribution Q_X when the true distribution is P_X and is given by

$$KLD(P_X||Q_X) = \sum_{x_i\in V_X} P_X(x_i)\log[\frac{P_X(x_i)}{Q_X(x_i)}] = E_X(\log[\frac{P_X}{Q_X}]) \quad (7)$$

where the expectation is with respect to the true density P_X. There are many powerful mathematical relationships that relate the information theoretic KLD directly to concepts and ideas from statistics. For example, the KLD is the expected log-likelihood ratio and the χ^2 statistic is twice the first term in the Taylor expansion of the KLD [21].

For random variables with continuous distributions, KLD is given by,

$$KLD(P_X||Q_X) = \int P_X(x)\log[\frac{P_X(x)}{Q_X(x)}]dx \quad (8)$$

where the integration is over the domain of X. The KLD is asymmetric (i.e., $KLD(P_X||Q_X) \neq KLD(Q_X||P_X)$ when $P_X \neq Q_X$), always takes non-negative values, additive, and is zero only if $P_X = Q_X$. To obtain a more symmetric measure, the Jensen–Shannon divergence (JSD) [22] is defined as

$$JSD(P_X, Q_X) = \frac{1}{2}KLD(P_X||M_X) + \frac{1}{2}KLD(Q_X||M_X) \quad (9)$$

where $M_X = \frac{1}{2}(P_X + Q_X)$. JSD is symmetric and bounded between 0 and 1.

2.4. Mutual Information

The mutual information of two random variables X and Y, denoted by $MI(X;Y)$ describes the information in probability distribution of one random variable X about the distribution of another random variable Y [2,20]. It is defined as the reduction in uncertainty in the random variable, say X after observing the other, Y, which can be shown to be equivalent to the relative entropy of the joint distribution of X and Y, $P(X,Y)$ relative to the product of their marginals, $P(X)P(Y)$:

$$MI(X;Y) = \sum_{x_i \in V_X} \sum_{y_i \in V_Y} P_{X,Y}(x_i, y_i) \log[\frac{P_{X,Y}(x_i, y_i)}{P_X(x_i) P_Y(y_i)}] = KLD(P_{X,Y} || P_X P_Y) \quad (10)$$

MI provides a measure of association or correlation between X and Y and is often denoted as "interaction information" between the two random variables. MI reflects the reduction in uncertainty (i.e., the information gained) for Y when X is known. Using the definition of entropy and conditional entropy, $MI(X,Y)$ can be expressed as

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (11)$$

Using the example of the SNP X and disease phenotype Y from above, MI between the X and Y can be easily calculated from the individual entropies $H(X)$, $H(Y)$ and the joint entropy $H(X,Y)$. $H(X)$ can be given by $H(X) = -\sum_{i \in \{0,1,2\}} P_X(x=i) \log[P_X(x=i)]$ with $P_X(x=i) = n_i/N$; $H(Y)$ is computed as $H(Y) = -\sum_{j \in \{0,1\}} P_Y(y=j) \log[P_Y(y=j)]$ with $P_Y(y=j) = c_j/N$. Here, n_i is the frequency of minor allele count i, c_j is the count of genomes with disease phenotype j and N is the total number of genomes. The joint entropy $H(X,Y)$ can be obtained as $H(X,Y) = -\sum_{i \in \{0,1,2\}} \sum_{j \in \{0,1\}} P_{X,Y}(x=i, y=j) \log[P_{X,Y}(x=i, y=j)]$. The joint distribution can be estimated as $P_{X,Y}(x=i, y=j) = n_{ij}/N$, where n_{ij} is the count of genomes with observed minor allele count i and disease phenotype j.

2.5. Interaction Information

The above definition of MI as a measure of interaction between two random variables can be extended to k variables using a multivariate generalized definition of interaction information also known as k-way interaction information (or KWII). It is defined as the amount of information (synergy or redundancy) that is present in the set of random variables, which is not present in any subset of these variables [23,24]. For a set of k random variables $S = \{X_1, X_2, \cdots X_k\}$, the KWII can be written succinctly as an alternating sum of the entropies of all possible subsets $\tau \subset S$ using the difference operator notation of Han [25]:

$$KWII(X_1, \cdots X_k) = -\sum_{\tau \in S} (-1)^{|S-\tau|} H(\tau) \quad (12)$$

As an aid to understanding the above equation, for the simple case of k = 3 discrete random variables, the KWII is given by

$$KWII(X_1;X_2;X_3) = -H(X_1) - H(X_2) - H(X_3) + H(X_1,X_2) + H(X_2,X_3) + H(X_1,X_3) - H(X_1,X_2,X_3) \quad (13)$$

The value of $KWII(S)$ can be both positive and negative where larger positive values indicate stronger interaction information (i.e., higher association) among the variables in S, while negative values indicate redundancy of information between the variables. As an example, for two SNP random variables X_1 and X_2 and disease phenotype random variable Y, the interaction information given by $KWII(X_1;X_2;Y)$ can be computed using entropies of all possible subsets as above. It can be interpreted as a measure of the synergistic association of the predictor variables X_1, X_2 with Y, i.e., how well X_1,X_2 explains Y—positive values will indicate that interaction between X_1 and X_2 enhances prediction of Y, while negative values will indicate that the prediction is reduced or inhibited.

3. Applications of Information Theory in Computational Biology

3.1. Gene Expression and Transcriptomics

Identifying, modeling and characterizing gene–gene relationships and their co-expression patterns is an important challenge in molecular and systems biology, as these dynamic interactions determine cellular phenotypes and represent causality inherent in developmental processes and regulatory pathways. Such associations are typically modeled computationally and statistically 'reconstructed' or 'reverse engineered' using Gene-regulatory Networks (GRNs, also referred to as Transcriptional Regulatory Networks) utilizing gene expression and mRNA abundance data obtained using technologies such as microarrays and RNAseq. More recently, large expression data availability from single cell experiments has spurred developments of new computational methods to infer GRNs, model regulatory interactions and gain insights into cell fate decisions and associated transcriptional state changes [26–30]. A GRN is modeled as an undirected graph or a network where the genes (and transcription factors) are represented as vertices (or nodes) and are connected by edges representing regulatory interactions between transcription factors and their targets. The gene expression data matrix (experimental or synthetic) typically provides the genes in the rows and samples and experimental conditions in the columns. The goal of the network inference method is to use the expression matrix to infer the set of regulatory interactions (direct or indirect regulation) between any two genes in the GRN, thereby predicting the edges in the network (Figure 1). Because mutual information (MI) provides the ability to capture non-linear dependencies between two variables, several methods for reconstructing GRNs use MI or associated information theoretic scores to infer the regulatory relationships. As a first step, these methods require the computation of the MI Matrix (MIM), a square matrix in which the element at the ith row and jth column is given by the MI between genes G_i and G_j. Often, this computational step has affordable complexity given that only pairs of MI computations based on bivariate probability distributions are needed to be estimated to complete the MIM. Several information theoretic methods have been developed to reconstruct GRNs over the past two decades [31–33]. For the remainder of this section, we will briefly review some of the popular information theory based methods that were discussed in-depth in some prior reviews [4,8] and then focus our discussion on more recent advancements in this area in the context of analysis of single-cell gene expression data.

One of the earliest approaches using MIM can be found in Relevance Networks [34] that simply links a pair of genes by an edge if the MI is larger than a given threshold. The approach has been introduced to infer relationships between RNA expressions and finding clusters of genes that affect cancer susceptibility to anticancer agents [35]. The Context Likelihood of Relatedness (CLR) algorithm [36] is an extension of the relevance network approach that calculates the statistical likelihood of a particular transcription factor/gene pair's MI value within its network context and filters many of the false (noisy) edges in the constructed GRN where one gene can weakly co-vary with many transcription factors simply by chance due to inherent noise in biological systems. It compares the MI between the transcription factor/gene pair to the background distribution of MI values for all possible pairs that include either the transcription factor or its target gene, and retains only interactions that have MI scores significantly above the background distribution. Both Relevance Networks and CLR suffer from a fundamental limitation—genes in indirect relationships separated by one or more intermediate genes may be highly co-regulated, resulting in numerous false positives. To address this, ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [37], a popular algorithm for reconstructing gene networks, was developed. ARACNE also starts with estimating the MIM between the genes, but it then filters spurious gene–gene associations using a two step process. First, each MI value between a gene pair is filtered using an appropriate threshold computed from the null-hypothesis of independence between the two genes, the null distribution being obtained by randomly shuffling the gene expression values for the given pair of genes. In the second step, ARACNE removes the vast majority of indirect associations using the data processing inequality (DPI) information theoretic

assertion that states that if there is an indirect association between two genes G_i and G_j through gene G_k, then $MI(G_i, G_j) \leq \min\{MI(G_i, G_k), MI(G_k, G_j)\}$. It means that gene G_i cannot have more information about gene G_j than gene G_k has about gene G_i and the post-processing cannot increase information. More recent extensions of ARACNE can be found in [38–40]. Time-Delay-ARACNE is proposed by Zoppoli et al. for GRN inference from time-course expression data [38], while higher order DPI is used to improve inference performance in hARACNE [39]. Improved computational performance is reported by Lachmann et al. in ARACNE-AP [40] by using an adaptive binning strategy to estimate MI.

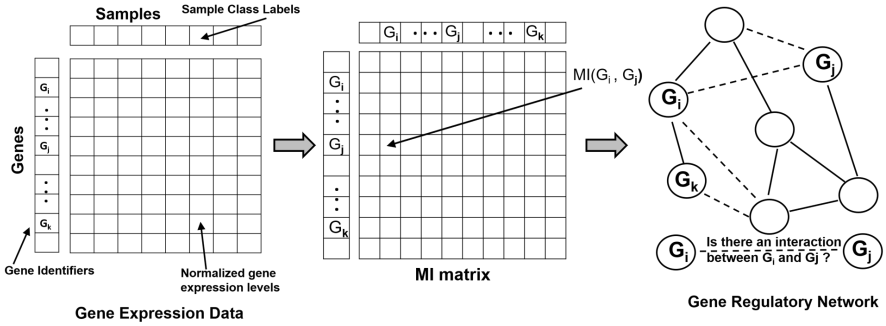

Figure 1. Schematic showing the Gene-regulatory Networks (GRN) reconstruction problem where undirected edges are inferred using information theoretic methods.

In the past few years, with the advancements in single-cell technologies for biological analysis, it has now become possible to quantify single cell transcriptomes in very large numbers. While that has made it possible to provide a finer-grained picture of the complex cellular processes and decipher heterogeneity inherent in gene expression across multiple tissue types, analysis of gene expression data from single cells comes with different data distribution patterns and characteristics (e.g., has a higher rate of zero values) often distinct from their bulk sample counterparts [8]. These present more complexities for statistical analysis (such as more technical noise and biological variability) [27]. As a result, GRN reconstruction methods developed for bulk sample expression data may not be suitable for data generated from single cells and may not take full advantage of the cell-to-cell variability to infer statistical relationships that can be used by information theory. Taking into consideration some of the different characteristics of single cell expression data, Chan et al. [27] developed a multivariate information theory based approach using partial information decomposition (PID) [41] computed from every triplet of genes in single-cell gene expression datasets. For three genes G_i, G_j, G_k, PID provides a decomposition of the information provided by two of the genes as source variables (say $S = \{G_i, G_j\}$) about the third gene G_k into three components—redundant, unique and synergystic information.

$$PID(S; G_k) = Synergy(G_k; G_i; G_j) + Unique_{G_j}(G_k; G_i) + Unique_{G_i}(G_k; G_j) + Redundancy(G_k; G_i; G_j) \quad (14)$$

The inference algorithm developed in [27] (named PIDC or "PID and Context") uses the Proportional Unique Contribution (PUC) metric defined as the average ratio of unique information between two genes G_i and G_j given the presence of a third gene G_k from the remaining genes S:

$$U_{G_i,G_j} = \sum_{G_k \in S \setminus \{G_i,G_j\}} \frac{Unique_{G_k}(G_i; G_j) + Unique_{G_k}(G_j; G_i)}{MI(G_i; G_j)} \quad (15)$$

Next, the confidence of an edge is calculated as the sum of the cumulative distribution functions of all PUC scores for each gene in the pair:

$$conf(G_i, G_j) = F_{G_i}(U_{G_i,G_j}) + F_{G_j}(U_{G_j,G_i}) \quad (16)$$

where $F_{G_i}(U)$ is the cumulative distribution function of all PUC scores involving gene G_i. Like CLR, estimating confidence of an edge by incorporating distributional properties of the PUC values for a pair of genes improves robustness to noise. It effectively identifies the most important interactions per gene, instead of relying on a simple ranking of the pairwise scores and choosing the best scoring edges across the whole network. Comparing PIDC with ARACNE and CLR, the authors reported marginal improvement in inferring the GRNs using simulated and publicly available data sets, attributing their success to larger sample sizes provided by single-cell data. However, it remains unclear as to how PIDC takes advantage of specific characteristics of single-cell data, such as its zero-rich nature.

Intrinsic noise in transcriptomic and gene expression data, owing to the stochastic nature of biochemical reactions driving the production of mRNAs and proteins within cells, makes GRN inference and reconstruction an extremely challenging problem. In addition to information theoretic methods, statistical methods such as Bayesian networks, tree based [42] and ordinary differential equation based [43] methods have been proposed to analyze gene expression data and reconstruct GRNs. An advantage of Bayesian networks and tree based methods is that they can learn directionality of the edges conjecturing if there is a directional influence from a regulatory gene to a target gene; information theoretic methods typically infer undirected edges. A head-to-head comparison of various methods to infer GRNs using single cell expression data has been presented recently in [8]. The methods were evaluated in their abilities to recover the correct set of GRN edges using ROC curves and Precision-Recall curves against reference networks from experimental assays, as well as in-silico reference networks from simulated data sets with single-cell characteristics. They reported that most of the assessed methods, including information theoretic methods, are not able to predict network structures from single cell expression data accurately and have differences with each other in the sets of identified edges. Also, their performances are inconsistent across data from different cell types and experimental conditions, as previously reported in another comparative study [44]. Both studies found no single best performing method across multiple types of data and simulation scenarios; rather, ensemble methods that average prediction using multiple approaches seem to work best [8,44]. The results from these studies emphasize the necessity to develop more accurate and optimized network modeling methods that are compatible with single cell expression data.

3.2. Alignment-Free Sequence Comparison

In this section, we discuss prominent applications of information theory for finding similarity between two genomic sequences without doing an actual alignment of the two genomes. Comparing two or more genomic sequences to find regions of similarities and dissimilarities and infer a measure of relatedness is vital for the success of basic phylogenetic and metagenomics research. Sequence comparison of genetic material across organisms allows discovery of gene structures and inferring their functional relationships based on the idea that higher similarity between sequences is a driver for higher structural and functional similarity and possible evolutionary relationships. While traditional methods for comparative sequence analysis and phylogeny reconstruction rely on computational algorithms for pairwise and multiple sequence alignments such as dynamic programming, they are often relatively slow in aligning two sequences, taking time proportional to the product of their lengths and are difficult to scale to whole genome comparisons. Furthermore, an underlying assumption supporting sequence alignment algorithms is the principle of co-linearity (homologous sequences have multiple linearly arranged and more or less conserved sequence stretches), which is often violated in biological genomes due to rearrangement events (e.g., in viral genomes) [45]. These and other challenges, such as a rapid drop in accuracy of sequence alignments when sequence identity falls below a certain critical point, have prompted development of alternative alignment-free methods for comparing large genomic sequences [46]. Among many different approaches to alignment-free sequence comparison (such as word frequency based methods inspired by linguistic analysis [47], graphical representation of DNA sequences [48,49], chaos game representation [50], iterated maps [51], spaced words [52,53]), applications of information theory have been key to developing methods utilizing the informational

limitations —> refined methods

content between sequences to be compared. In the remainder of this section, we briefly review some of the key articles using information theory for alignment-free sequence comparison.

In the feature frequency profiles (FFP) method [54] developed by Sims et al., the authors proposed to estimate the distances between nucleotide or amino acid sequences using a two step process. In the first step, fixed length words (or k-mers) and their frequencies are extracted from each genomic sequence by sliding a window along the genome. Normalizing each frequency with the sum of all k-mer frequencies generates a probability distribution for the genome. This process leads to the conversion of each genomic sequence into its FFP, F_k, represented by the distribution of each word in the genome. In the second step, pairwise distance between two sequences i and j can be calculated using the Jensen–Shannon divergence (JSD) between their respective FFPs, F_i and F_j. The JSD metric symmetricizes the asymmetric KLD distance metric between two probability distributions. The JSD serves as a measure of distance between the two genomes and can be used to generate phylogenetic trees from multiple genomes. The authors also used information theory to address a key question in their approach—how can one choose an optimal value of k (the word length). Choosing a large value of k will run into higher computational complexity and possibly unreliable estimates of k-mer probability distributions. Choosing a smaller value will result in words that are too small, commonly occurring in all genomes and does not have the information to distinguish between the genomes. To address this, the authors computed lower and upper bounds of word size k. The lower limit is given by the word length that is most frequent in a genome of interest. The upper limit is derived using the $k-2$ Markov model, which says that one can predict the frequency of a word of length k using the frequencies of its $k-1$ and $k-2$ subwords. So for a given word length k, the expected FFP, $\hat{F}_k$, can be computed and KLD or relative entropy is then calculated between F_k and $\hat{F}_k$ for all values of k from one to infinity and summed to get CRE (cumulative relative frequency). The CRE represents the accuracy of predicting FFPs for all lengths $\geq k$, given the prior distributions F_{k-1} and F_{k-2}. The value of k at which the CRE approaches 0 becomes the upper limit of word size for use in genome comparison.

More recently, probabilistic data structures for k-mer counting have been used to enable alignment-free genome comparison [55,56]. A critical advantage of these data structures is higher memory efficiency allowing them to be used for comparing many sample sequences at the same time using next generation sequence data. One of these methods, kWIP, hashes all k-mers from a genomic sequence i into a probabilistic data structure called a Sketch [56,57], S_i, that are numeric vectors of fixed size for every sequence. Given a set of many sequences (e.g., from a metagenomics experiment), genetic relatedness between any two sequences i and j is then calculated by computing an inner product between the two sketches S_i and S_j, weighted by their informational entropy across the population set. This procedure down-weights uninformative k-mers (highly abundant or present in very few samples) and produces a kernel matrix, K, of pairwise inner products that is then normalized using the Euclidean norm $K'_{ij} = K_{ij}/(K_{ii}K_{jj})$ and converted to distance matrix $D_{ij} = \sqrt{K'_{ii} + K'_{jj} - 2K'_{ij}}$ containing the genetic distances between every pair of sequences in the population. A schematic illustrating the FFP and kWIP approaches is presented in Figure 2.

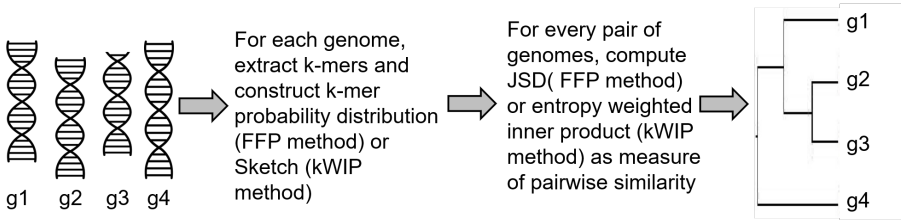

Figure 2. Schematic showing an overview of alignment-free sequence comparison.

Over the past few years, many alignment-free algorithms have been proposed [58–60] to solve previously intractable challenges in phylogenomics, many of which have been compared in [9,45] using data from three categories—gene tree inference, genome-based phylogeny and horizontal gene

transfer. Assessing the degree of topological disagreement between the inferred and reference trees, the report found that no single method performed best across all the data sets. When compared using eight data sets from the whole-genome phylogeny and horizontal gene transfer categories, FFP was found to be among the top five best performing methods out of 55 variants of 11 alignment-free phylogeny tools tested. These suggest that improved information theoretic distance metrics to assess similarities between genomes need to be developed as they were often more challenged in deciphering finer complex organization levels in the sequences [45]. With genomics continuing to lead in terms of data growth and availability through production of high-throughput sequencing data, although information theory based methods are observed to be relatively more memory efficient and computationally inexpensive than other methods for smaller data sets, bigger data sets pose more serious challenges, for example, substantial memory overhead when using longer k-mers. Analyses of very large data sets containing beyond tens of thousands of samples will benefit from more optimized implementations of methods such as kWIP that performs pairwise computations, including parallelization across distributed processing systems. Therefore, it remains a challenge in alignment-free phylogeny to develop both novel metrics and efficient implementations that can scale to handle larger data sets.

3.3. Sequencing and Error Correction

Information theory based approaches have notable applications in genome sequencing. DNA sequencing technologies have continuously evolved over the past two decades, leading to a substantial increase in throughput and decrease in cost. Current technologies are mainly based on a two-step approach. First, fragments of the target sequence/genome are read via shotgun sequencing. Then, these fragments, so-called reads, are assembled to reconstruct the original sequence. This reconstruction task can be seen as a problem of assembling a string from its substrings, where three aspects play an important role in the completion of the task: length of the fragments; number of reads per base (coverage depth); and fragment error rates. In the case of DNA sequencing, these aspects are highly dependent on the sequencing platform. Average read length typically ranges from 100 to 10,000 base pairs (bp), although platforms such as PacBio and Oxford Nanopore Technologies MinION can reach maximum read lengths in the order of 60–100 kbp [61,62]. The number of reads typically ranges from 100 s to 1000 s of millions. And the error rate can range from 0.1% to 20% [63,64].

Given a set of reads, the most crucial challenge for the genome reconstruction task is to determine whether complete reconstruction of the target sequence is possible from the given set of reads. Additionally, other important challenges include finding the minimum fragment length and coverage depth needed for the complete reconstruction with a given reliability, evaluating the impact of the read error rates on the assembly performance and on the requirements in terms of read length and coverage depth, and designing ways to compare different sequencing technologies and assembly methods. These challenges have been recently tackled from an information-theoretic perspective by investigating the fundamental limits of genome reconstruction. From an information theory point of view, a sequencing task can be seen as a decoding task where the genetic information is first encoded in terms of nucleotide bases and then transmitted during sequencing through a noisy channel in form of reads (Figure 3).

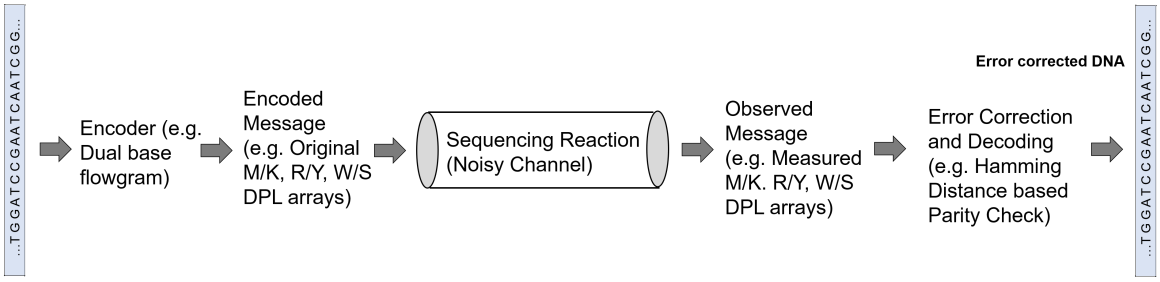

Figure 3. Illustration of an information theoretic communication model inspired representation for error correction in genomic sequencing based on the work of Chen et al. [65].

Motahari et al. [66] investigated the fundamental limits for the read length and the number of reads needed to allow complete sequence reconstruction, assuming error-free reads and DNA sequence modeled as an i.i.d (independent and identically distributed) random process. They calculated the minimum read length required for complete sequence reconstruction based on the Rényi entropy [67] of order two, which is defined as $H_2(p) = -\log_2 \sum_i p_i^2$, where p is the probability distribution on the alphabet {A,C,G,T} and i iterates over each of these values. More specifically, given read length L and original sequence length G, they calculated the normalized read length $\overline{L} = L/\log_2(G)$ and showed that if $\overline{L} < 2/H_2(p)$, complete sequence reconstruction is impossible, regardless of the number of reads. If read length is above that threshold, then complete reconstruction requires that enough reads exist to cover every base least once. These findings were later extended to consider scenarios that do not require the i.i.d DNA sequence assumption or reads to be error-free [68–70].

Gabrys et al. [71] studied these questions in the more general context of string reconstruction, motivated by applications in DNA-based data storage systems [72], discussed later in this section. They investigated the fundamental limits of read overlap and their impact on the read length requirements, and established the minimum Hamming distance between reads to allow proper sequence reconstruction. Marcovich and Yaakobi [73] followed the work in [71], by exploring two models for string reconstruction: one in which not all substrings are received, and another in which all substrings are received, but with errors. They also proposed substring constraints based on the Hamming distance between reads.

In another proposed method [74], Si et al. investigated the closely related haplotype assembly problem, where the goal is to reconstruct haplotype sequences (i.e., string of single nucleotide polymorphisms (SNPs) on a single chromosome in a homologous pair). Using an information-theoretic perspective, they formulated this goal to be the successful recovery of two sources of information being communicated through a channel: the haplotype information and the chromosome membership. They studied the required conditions to allow a reliable reconstruction with and without error-free reads and showed that the requirements in terms of number of reads for the erroneous case is of the same order as in the error-free case. These limits were defined using the Fano's inequality [2].

More recently, in a novel approach, Chen et al. [65] combined sequencing-by-synthesis with an information theory-based error-correction algorithm, as outlined in Figure 3. They proposed a communication channel with three rounds of fluorogenic DNA sequencing [75] and use information theory principles to analyze information redundancy. In each round, the sequencer generates a dual-base array, which is combined in a later stage to infer the four-base DNA sequence. For example, one of the rounds generates the so-called *MK* dual-base array to determine whether the DNA base is *A* or *C* (encoded as *M*) or *G* or *T* (encoded as *K*) for every position of the sequence; in this case, *K* and *M* are called degenerate bases. Considering a random DNA sequence, the entropy of a sequence of length L is $2L$ bits, while the entropy of a dual-base array is L bits, assuming no sequencing error. Therefore, if two different dual-base arrays are used, the original sequence can be fully reconstructed. However, as the actual entropy of each dual-base array is lower than L, because of sequencing errors, a communication channel based on only two rounds, one for each dual-base coding, would be prone to propagate sequencing errors. Based on the experimental error rate of fluorogenic DNA sequencing, the authors showed that using three dual-base array provides enough information for sequence reconstruction with error correction. Besides the *MK* dual-base array, they used the *RY* array, where *R* means *A* or *G* and *Y* is *C* or *T*, and the *WS* array, where *W* means *A* or *T* and *S* is *C* or *G*. By combining the information encoded by the three degenerate arrays, the decoder can correct an error of one of the rounds. Experimental results in [65] showed that this strategy reduces the error rate in a 250 bp sequence from 0.96% to 0.33%. In a recent study, Mitchell et al. [76] benchmarked several computational error-correction methods for sequencing data and reported that Chen et al.'s method [65] suffered from increased computational cost compared with other methods that limits its scalability to larger data sets.

The same concept of degenerate bases is used in [77,78], but with the focus of increasing information capacity for DNA based data storage when storing results for multiples reads from

a target sequence. By using different encoding bases (original and degenerate ones) to encode the observed base frequencies for each position of the sequenced reads, the need to store all reads is eliminated. As a finite alphabet of encoding bases, including pure and degenerate bases, the observed base frequency will most likely not match any of the bases in the alphabet. Anavy et al. [77] solved this problem by using KLD to make the best approximation, while Choi et al. [78] used a clustering approach. They both used Reed–Solomon coding [79] as a correction mechanism.

Improvements in error correction for long-read sequencing (LRS) technologies, such as PacBio and MinION, is one of the current challenges in DNA sequencing [80,81]. LRS-based approaches have shown advantages in sequencing complex regions, such as those with extreme GC-rich, long tandem repeats, and interspersed repeats, when compared to prevailing short-read approaches [82,83]. However, their high error rates in DNA sequencing are still considered a main drawback [84], and information theoretic methods have the capability to reduce it, as shown in [65]. Information theory offers the potential to further contribute in this area by defining fundamental limits for the coverage depth of DNA reads, to guarantee full reconstruction, error correction, and minimum redundant information. Using information theory for biology inspired and error-free information storage is another emerging area of research [72,85–87]. It has drawn more attention lately because of the novel use of DNA as an archival medium and holds a lot of promise for future research to satisfy the exponentially increasing demand for information storage.

3.4. Genome-Wide Disease-Gene Association Mapping

Over the last two decades, fueled by studies and reference projects of human genetic diversity and developments in high-throughput sequencing technologies, genome-wide association studies (GWAS) have been the primary tool to investigate and establish the connection between the genetic data and the observed characteristics or phenotypes, such as occurrence of human diseases or disease biomarkers. For the human genome, the most common form of observed genetic variations are bi-allelic single nucleotide polymorphisms (SNPs) spread across the human genome. The data is usually represented as a numeric matrix where the rows are the genomes from multiple individuals and the columns represent the SNPs. Each bi-allelic SNP (say the ith SNP in the data) can be modeled as a discrete random variable (say, X_i) taking values from the set $\{0, 1, 2\}$, representing the number of copies of the minor allele in the genome at that position. Also for every individual sample, there is an observed disease phenotype (say, modeled as random variable Y), which can be discrete in the simplest case (e.g., presence or absence of a disease), or it can be a continuous valued phenotype (e.g., blood pressure measurements or blood cholesterol values).

Many information theory based approaches have been proposed for disease-gene association studies in the last twenty years. Following a similar characterization as the one proposed by Wu et al. [88], these approaches can be based on: (1) single SNPs; (2) haplotypes; (3) genes; and (4) gene–gene (GXG) interactions. Single SNP based association methods (also known as single-locus tests) are the most common and simplest approaches to identify the disease associated SNPs. Two examples of association tests based on entropy for single SNP-based analysis can be found in [89,90]. These methods devise test statistics treating every SNP as an independent random variable and have low computational complexity proportional linearly to the number of SNPs across the genome considered for the study. However, single SNP based methods suffer from low power to identify causal associations, especially when the SNPs involved are spread across multiples genes [88]. Furthermore, for many such multi-factorial diseases, pathogenic genetic variants usually have a low population of minor allele frequencies (MAFs), making it difficult to identify their effects for relatively small sample sizes [91]. Haplotype-based approaches look at combinations of marker alleles that are closely linked on the same chromosome and tend to be inherited together as a unit, instead of alleles from individual SNPs. This characteristic makes these approaches more suitable to analyze complex multi-factorial disease phenotypes than the single SNP-based methods [88]. Entropy-based tests using haplotypes can be found in [92–94]. Their results showed that these tests tend to outperform non-information theoretic

tests. However, as pointed out in [88], their main drawback is the computational cost associated with inferring the haplotype phase and frequencies needed for the tests. Closely related to haplotype-based methods are gene-centric approaches that consider genic variants within one gene as testing units and are specially powerful when there is more than one disease variant in a gene. These approaches are specially advantageous when samples come from non-homogenous populations. An entropy-based association test for gene-centric analysis can be found in [95]. In addition to the association test, they also proposed a penalized entropy measure that is used to cluster genotypes and decrease the degrees of freedom of the association test. Gene–gene interaction studies comprise another prominent direction of investigative strategies, those that involve devising test statistics specifically geared towards studying complex, multi-factorial diseases. Many of these proposed methods, discussed below, use information theoretic metrics designed based on Shannon's entropy. The goal of these information theory based methods is to develop test statistics or scores that measure gene–gene (GxG) statistical interactions (also known as epistasis, [96]) between a set of SNPs and the disease phenotype, and also identify statistical interactions between the genotype and environment (GxE) when environmental variables (factors external to the genome) are included in the study design (Figure 4).

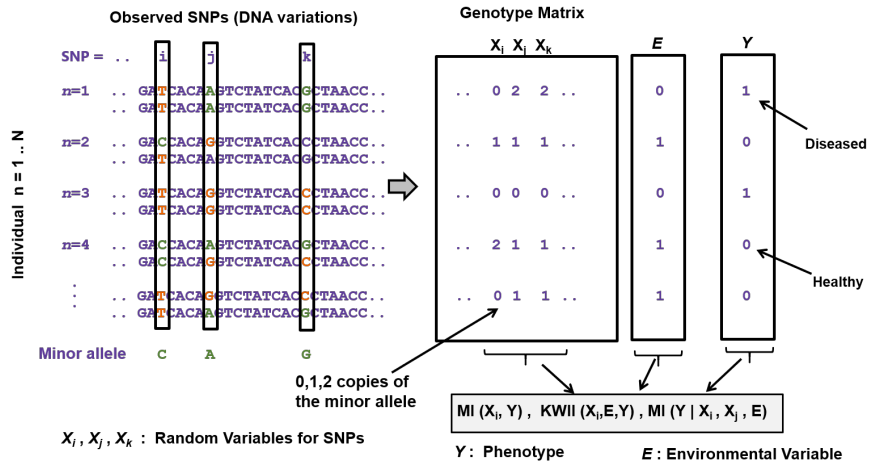

Figure 4. Overview of the genome-disease association problem and some example information theoretic metrics calculated for identifying GxG and GxE, a binary phenotype and environmental variable is shown for simplicity.

Some of the early information theory based approaches focused on detecting interactions by proposing test statistics based on MI involving two SNPs and the disease phenotype from case-control studies in which a group exhibiting the phenotype/disease is compared to a control group. For example, Fan et al. [97] computed MI between two genetic variants X_i and X_j separately in cases and controls and used their difference $\mathrm{MI}_{cases}(X_i; X_j) - \mathrm{MI}_{controls}(X_i; X_j)$ as the test statistic to estimate information gain. The idea is to use the MI between the SNPs in the control group as an approximation of the MI in the general population and identify pairs of SNPs with effects substantially bigger than that among the cases. In another proposed method by Yee et al. [98], the difference between the entropy of the phenotype,$H(Y)$, and the conditional entropy of the phenotype given a pair of genetic variants, $H(Y|X_i, X_j)$, is normalized with regard to the overall entropy of the phenotype to get the test statistic $[H(Y) - H(Y|X_i, X_j)]/H(Y)$ that normalizes the information gain with regard to the overall observed entropy of the phenotype. This is also referred to as normalized MI that quantifies the proportion of information contained in the interacting SNPS, X_i and X_j, influencing the phenotype Y. In another method, Dong et al. [99] have proposed the ESNP2 (entropy-based SNP–SNP interaction) method integrating two-locus genetic models. Their test statistic is defined as

$$\triangle R_{i,j} = \frac{\min(H(Y|X_i), H(Y|X_j)) - H(Y|X_i, X_j)}{\min(H(Y|X_i), H(Y|X_j))} \tag{17}$$

and uses conditional entropy of the phenotype random variable $H(Y|.)$ to measure interaction effects of SNPs X_i and X_j whenever marginal effects exist. All these methods rely on variations of mutual information (MI) and conditional-MI based two-SNP test metrics and heuristic algorithms of mining the genome looking for two-SNP combinations showing significantly high association scores. A detailed systematic review of these methods, distributional properties of these metrics and their mathematical derivations, and associated algorithms using these information-gain-type quantities for feature selection in the context of genetic analyses can be found in [100].

While the methods summarized above investigate GXG interactions considering relatively simpler two-way SNP–SNP interactions, evidence suggests that higher-order multivariate GxG interactions can contribute to a number of complex traits [101]. Addressing this, several methods have been proposed using more intricate information theory based models that consider the contribution of more than two SNPs to the onset of a phenotype [102–113]. For such models, information gain through higher order (with more than two random variables) GxG interactions in explaining the phenotype states as more SNPs are added to a model can be defined from a synergy/redundancy point of view [24,103–106,108] that quantifies the effects of two or more SNPs compared to their individual effects. The synergy can be positive, when the joint effect of multiple SNPs is larger than the sum of the individual single SNP effects, or negative, when the joint effect is smaller than the sum, indicating information redundancy among SNPs. Some of the prominent methods addressing higher order interactions through synergy and redundancy use multivariate generalizations of MI such as *K*-way Interaction Information (KWII) to parsimoniously model both GXG and GxE interactions and help gain deeper insights into the underlying disease causation pathways by combining them in a single analytical framework [114–120]. All these methods depend on detection of higher order interactions in terms of synergy/redundancy that relies on robust empirical estimations of metrics such as entropy, MI and KWII. A study by Sucheston et al. [121] systematically evaluated information theoretic metrics such as KWII along with established non-parametric methods, such as Multifactor Dimensionality Reduction [122] and Restricted Partitioning Method [123], comparing power and Type I error. They found that information theoretic models have more flexibility and have excellent power to detect GxG interactions under a variety of conditions including genetic heterogeneity that characterize complex diseases.

Building on similar ideas, epistasis networks proposed by Moore et al. [124] provided a way of using GxG combined with computational network theory. Using the SNPs and their relationships, a network can be created where each node is an SNP and each edge is a KWII between a pair of SNPs in the presence of an observable trait or phenotype, such as bladder cancer susceptibility. An edge is included in the network if the KWII strength is above a threshold that is determined using permutations of the SNP genotypes and phenotype. Once the epistasis network is created, the authors performed network analysis to reveal interesting information. For example, the degree distribution of the vertices in the network was found to closely follow the power law distribution, and hence the network was approximately scale free [125].

In terms of target phenotypes, most of the work described above considers dichotomous phenotypes for case-control studies. Alternative strategies to identify GxE interactions are the case-only and the family-based designs. Case-only studies [126] are used to identify interactions using data from only affected individuals as used in [127,128]. For example, Kang et al. [127] used case-only design to devise a multi-SNP test statistic $2N(1 - H(Y))\log(W)$ that is shown to be asymptotically χ^2 distributed under the null hypothesis of no association with $W - 1$ degrees of freedom. For a k-SNP loci, $H(Y)$ refers to the entropy of the phenotype Y obtained using the counts of each unique genotype observed across all the k SNPs; W is the total number of genotype combinations observed on these loci. These tests are shown to have higher power when compared to the standard χ^2 statistical association test. Family-based studies take into account the family relationship of the genomic sequences when comparing them. It considers, for example, that an allele associated with a disease will be transmitted to the affected offspring more often than that expected by chance—the Transmission Disequilibrium Test (TDT) [129]. Examples of such disease association test metrics for family-based study designs can

be found in [88,102,130]. Notably, Zhao et al. [130] designed a novel TDT statistic using entropy to generalize the original TDT statistic and incorporated non-linearity in its definition. They demonstrated that the entropy-based TDT test is more powerful than the original TDT test. Besides analyzing binary traits, often GWAS needs to detect gene-disease associations for quantitative traits involving non-discrete real values. Information theoretic methods dealing with such quantitative phenotypes and environmental variables can be found in [116,117,131] that use differential or cross-entropy based generalizations of multivariate test statistics. In a recent review, Galas et al. [132] presented a detailed discussion on the information theoretic formalism for gene association with quantitative phenotypes.

Recently, Tahmasebi et al. used an information theoretic approach to investigate the fundamental limits of GWAS parameters [133]. Their study proposed an abstract probabilistic model to detect the causal subsequence of length L for a specific phenotype using a dataset of N individuals with genomes of length G and their observed characteristics, where the presence of external environmental factors make the relationship between the causal subsequence and the observable characteristics a stochastic function. With increasing value of the model parameters N, G and L, the authors reported observing a threshold effect at $(G/N)H(L/G)$ (H denoting binary entropy) that was then used to formulate the capacity of recovering the causal subsequence using information theory. This idea was further extended in a subsequent article [134] by comparing mixed and unmixed populations.

With increasing availability of high-density SNP data from next-generation sequencing techniques like genotyping-by-sequencing, methods for identifying GxG and GxE interactions have rapidly evolved to be an essential technology for association studies. While regression based methods and Bayesian statistics [135,136] have been the primary workhorses for GWAS, information theory has played a crucial role in advancing the field and has been often combined with statistical approaches in reducing dimensionality and devising novel solutions. Nevertheless, there are some open questions that need to be addressed in future research. Although many definitions of statistical interactions and associated test statistics are proposed using information theory, the field would benefit from a unified definition and interpretation of statistical interactions. Additionally, often the underlying distributions of the test statistics under the null and the alternative hypothesis are unknown and more studies are needed as in [137], focusing on investigating the asymptotic behaviors of the estimators involved. Computing higher order test statistics like $KWII$ is computationally expensive as it necessitates entropy computations of all possible subsets of SNP combinations—for example, computing KWII for a set of two SNPs and a disease phenotype entails $2^3 - 1$ entropy computations with three random variables. As a result, computing all possible k-SNP combinations in a GWAS study can quickly become computationally infeasible for larger values of k. This also increases the number of tests and can reduce power if multiple-testing corrections, such as Bonferroni correction [138], are applied. As a possible remedy, often single SNP analysis is used as pruning step to reduce the number of SNPs to be analyzed prior to interaction analysis, and test significances are estimated through efficient permutation strategies [139]. Another challenge is that the information theoretic test statistics and estimators are, in many instances, still to be adapted to address practical challenges in genetic studies such as accounting for missing genotypes, genotyping errors, phenocopies and genetic heterogeneity. Finally, although some studies exist comparing computational methods for epistatic interaction detection (e.g., [140]), research is still lacking with respect to well-defined studies comparing the powers and performances of several key information theory based as well as other statistical approaches under a variety of experimental conditions. The field will immensely benefit from such systematic studies in the near future.

3.5. Protein Sequence, Structure and Interaction Analysis

Proteins are key macromolecules, which mediate their function by interacting with other molecules, including other proteins, nucleic acids, metabolites, and lipids. Major biological processes, such as immunity, metabolism, signaling, gene expression, and molecular machines are controlled through protein interactions [141–143]. Due to the critical nature of protein interactions, a key way of investigating

an unknown protein is to determine what interactions occur between proteins. The experimental determination of protein–protein interactions (PPIs), using methods such as yeast-two-hybrid, tandem affinity purification, mass spectrometry, DNA and protein microarrays [144–146], is time consuming and relatively expensive, and hence, there is a need for accurate computational methods for determining interactions among proteins. With the availability of genomics and proteomics data, several computational methods have been proposed for determination of protein–protein interaction partners. These examples cover different approaches, including Bayesian methods [147], graph-based methods [148], phylogenetics-based methods[149,150], and information theoretic approaches. Applying information theory to protein sequence analysis allows researchers to identify these interactions more easily and without the cost associated with experimental methods. For protein structural analysis, information theoretic methods also offer the advantage of being able to capture relationships at the level of protein families, in contrast to other computational techniques such as PPI structural modeling or docking, which typically operate at the level of individual protein complexes. This is because, information theoretic techniques for protein interaction modeling can capture dependency information across multiple sequences to calculate useful metrics, and are thus able to discern interactions at the protein family level.

An information theoretic approach to infer PPIs is based on the concept of coevolution. The key biological concept that underpins this approach is that interacting proteins adapt together due to evolutionary pressures, which gives rise to relationships that can be observed while observing a large set of data that spans this evolutionary history [151] (such as observing similar gene sequences or gene expression profiles from different cell types or organisms [152]). These relationships can be identified by applying information theoretic concepts such as mutual information (MI). As previously described, MI measures the degree to which random variables are co-dependent. From a multiple sequence alignment of several related proteins, the amino acid distribution at every site can be thought of as a random variable. Following this, MI can quantify the degree of reduction of uncertainty at one site, based on the knowledge of the distribution at a different site. In a notable initial application of these approaches, Giraud et al. [153] applied the concept of coevolution to proteins in their study in 1998. Since that time, several attempts have been made to refine this measure for their applicability to proteins. Several studies [154–156] have described approaches to model the effect of phylogenetic relationships between the proteins and separated this from the coevolution between pairs of sites within a multiple sequence alignment, and have demonstrated that these techniques are able to improve the applicability of such metrics to identifying PPIs. Figure 5 shows the role of information theory in a typical coevolution approach.

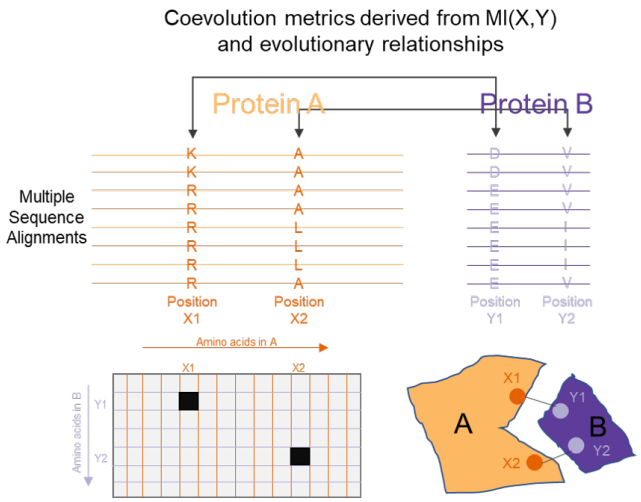

Figure 5. Simplified illustration of how mutual information (MI) can be used to capture links between two proteins from their Multiple sequence Alignments. High MI between 2 columns of the alignments are associated with interactions in their structures.

These methods also have a significant impact in the field of structural biology. A key element to understanding and predicting the function of a protein is its structure. While a structure is largely determined by its sequence, over the years, structures for several proteins have been determined by X-ray crystallography, Nuclear Magnetic Resonance (NMR) methods, and more recently Cryo-Electron Microscopy (CryoEM). As the number of structures solved and deposited in the Protein Data Bank (PDB) keeps increasing, the vast explosion in data has allowed the application of information theoretic approaches to play impactful roles in structural biology [157].

Since complementarity is a key feature of three-dimensional protein interactions, coevolution would suggest that when a substitution occurs, a compensatory substitution would occur in an interacting partner, that would be captured by the metric regardless of where in the sequence the mutation occurred. Validation of coevolution-based approaches discussed previously has been performed by studying the structures. Specific examples encompass a variety of applications, some of which have been described in a study by Little et al. [10]. This cross-functional work includes a study of coevolution in the PDZ domain of human protein Erbin, also among related, previously known and annotated functional domains from the PFAM database, as well as across known catalytic sites. Other studies have also demonstrated the immense potential that coevolution offers to structural and systems biology. Specialized approaches for modeling protein coevolution are required since amino acids have several interacting partners, and MI based metrics capture interactions that may be due to indirect coupling [157]. This additional complexity is resolved in methods such as DCA[158] or GREMLIN[159], which are able to highlight direct interactions and remove the noise from indirect interactions.

In a recent study, Cong et al. [160] have applied these techniques to large-scale data from proteomics studies. This study investigated coevolution between 5.4 million pairs of proteins in *Escherichia coli* and between 3.9 million pairs in *Mycobacterium tuberculosis*, and reported that in conjunction with structural modeling, they were able to predict PPIs with an accuracy much higher than that found by traditional proteome-wide two-hybrid screens. This study only highlights the potential for information theory methods to go beyond traditional methods and help identify PPIs for unexplored organisms.

Not addressed here, but other equally challenging and interesting uses of information theory to analyze PPIs are addressed in a review article by Mousavian et al. [5]. In that review, the authors discuss in detail applications of information theory in dealing with PPI networks, describing a variety of applications such as protein complex identification, network complexity analysis, finding subnetwork markers, and other questions addressed by previous studies applying information theory for PPIs. In contrast, for this review, we have highlighted some studies that build on the concept of sequence coevolution and discuss the future scope of these methods in solving unsolved problems. As these methods are further refined and integrated with structural biology, as in Cong et al. [160], they can be further improved with the goal to map all cellular PPIs, enabling discovery of the "interactome"—the cellular protein interaction network. This represents a challenging task of evolving from the current paradigm of discovering interactions between specific proteins or protein families to mapping at the scale of millions of proteins. Other exciting and emerging applications in the structural domain include recent studies of specificity in these interactions and how they can help guide in-silico assemblies of protein complexes [157].

3.6. Metabolic Networks and Metabolomics

Metabolic networks are networks that describe the physiological and biochemical properties of a cell. The components are metabolites, chemical reactions that transform the metabolites to each other catalyzed by metabolic enzymes, metabolic pathways, as well as the regulatory interactions that guide these reactions. Unlike other types of systems biology networks, studies of metabolic networks have been limited due to difficulties in generating large amounts of metabolite concentration data and lack of knowledge about the kinetic properties of the involved reactions. The recent rapid evolution of metabolomics based on mass spectrometry and nuclear magnetic resonance has made possible faster and more in-depth studies of metabolic networks. In this section, we will introduce some studies that

have applied the concepts of information theory in understanding metabolic networks. An overview illustrating some of the approaches discussed is presented in Figure 6.

Reconstruction of metabolic networks remains a central topic similar to other types of biological networks. A top-down approach to metabolic network reconstruction is by reverse engineering of metabolome data. Popular methods for inference of regulatory networks, ARACNE [37] and CLR [36] have been applied to reconstruct cellular metabolic networks [161]. Both these methods leverage mutual information (MI). A method based on conditional MI has also been used [162]. These methods can detect non-linear correlations compared to simpler correlation-based relatedness scores. Saccenti et al. [163] proposed a "wisdom of crowds" approach that considers the consensus obtained from four different approaches, ARACNE, CLR, PCLRC (Probabilistic Context Likelihood of Relatedness on Correlations) [164] and Pearson correlation.

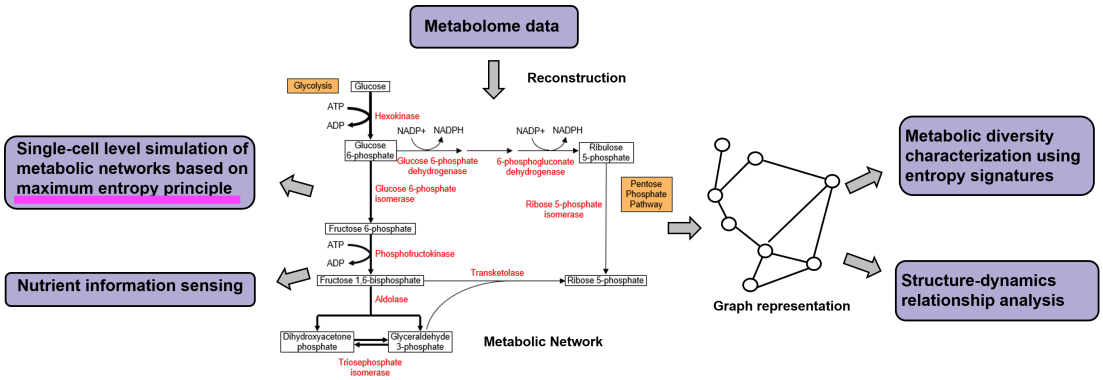

Figure 6. An example of metabolic networks. Nodes represent metabolites and are transformed into each other through chemical reactions catalyzed by metabolic enzymes (red). Metabolic pathways are formed by a linked series of chemical reactions that collectively perform a biological function. Reconstruction of metabolic networks is typically done by reverse engineering of metabolome data, where information theoretic methods ARACNE and CLR have been applied. Reaction fluxes depicted by the width of the reaction arrows can be predicted using FBA. Also shown are examples of downstream analysis of a metabolic network using information theory for metabolic diversity characterization, structure-dynamic relationship analysis, single-cell level simulation of metabolic networks with FBA and nutrient information sensing.

Molecular networks built from metabolite profiling can exhibit a large degree of diversity across individuals and this variability reflects the intrinsic diversity observed among the individual metabolic phenotypes. To characterize this diversity, the concept of entropy can be used [163]. Importantly, both entropy profiles of single metabolites and entropies of one metabolite relative to others that characterize dependencies and correlations between metabolites in a network context need to be considered. In the recent review published by Everett et al. [165], metabolic entropy is considered as one of the four fundamental approaches to the generation and utilization of metabotype data for metabolic phenotyping in diagnosis and prognosis, another being the metabolic network itself.

An important goal of metabolic network studies is understanding the dynamical behaviors of metabolic networks and the functions generated by them. Despite their topological complexity, metabolic networks avoid complex dynamics and maintain a steady-state behavior. A combination of the Shannon entropy and the word entropy [166] capable of separating different dynamic regimes in metabolic networks have been used to reveal that this pronounced regularization of dynamics is encoded in the network topology [167]. Nykter et al. [168] also studied network structure-dynamics relationships, using Kolmogorov complexity as a measurement of information distance between pairs of network structures and between their associated dynamic state trajectories. The possible dynamics of metabolic networks was studied in Grimbs et al. [169] by using a kinetic model. The enzyme kinetic parameter space was sampled and the metabolic dynamic states were evaluated statistically. MI was then used as one of the three distinct measures to assess the relative impact of kinetic parameters on the

stability and robustness of metabolic networks. Cellular metabolic systems form self-assembled aggregates and the activities of cellular enzymes can also exhibit spontaneous spatial-temporal functional structures. Entropy is a useful concept in the study of these dynamical systems [11]. In particular, Kolmogorov–Sinai entropy, which can be estimated from a finite number of observations using a family of statistics named Approximate Entropy (ApEn), provides a good measure of the complexity and information for the study of attractors in biochemical systems.

Metabolic networks are often modeled and simulated using Flux Balance Analysis (FBA) in order to study the physiology of the relevant microorganism or cell. FBA can predict metabolic reaction rates, also known as fluxes, without using kinetic parameters, by representing a metabolic network in the form of a set of mass balance equations based on the stoichiometry of each reaction, and computing reaction fluxes to match biomass production rate to a measured growth rate. However, substantial cell-to-cell growth rate fluctuations exist even in well-controlled steady-state conditions. De Martino et al. [170] introduced a generalization of FBA to the single-cell level based on the maximum entropy principle [171]. The idea is to look for an as-random-as-possible distribution over fluxes that is matching the experimentally measured average growth rate. This maximum entropy metabolic modeling has been shown to provide a better match to experimentally measured fluxes and it makes a wide range of predictions such as on flux variability, regulation, and correlations.

Finally, metabolic networks transform nutrients into biomass and it is of interest to understand how a cell acquires information on nutrient availability through nutrient sensing and how a metabolic network uses this information. Wagner at al. [172] proposed a way to relate the nutrient information to the error in cell's measurement of nutrient concentration in the environment. FBA was then used to show that nutrient sensing inaccuracy is translated logarithmically into reduced cell growth, and for microbes like yeast, cells would need to estimate nutrient concentrations to very high accuracy to ensure optimal growth.

3.7. Connections to Optimization and Dimensionality Reduction in Biology

In this section, we briefly review some key applications connecting information theory with optimization, notably through Maximum Entropy Production (MEP) and then discuss information theory in the context of dimensionality reduction for omics analysis.

3.7.1. Optimization in Biology

Natural selection can be considered an optimization process itself, given environmental constraints, leading to "survival of the fittest" . While mathematical models have been used to model biological processes ranging from microbial metabolism [173,174] to plant growth [175,176] to human disease [177,178], Maximum Entropy Production (MEP) has served as an ultimate measure of fitness by "survival of the likeliest" [179,180]. MEP is defined by [180] as a memoryfull (or path-dependent) change in a system's state probability distribution. While a memoryless entropy change between distributions P and Q is often measured by Kullback–Leibler Divergence (KLD) in physics, MEP, as applied to a biological system, reflects the thermodynamic likelihood of all chemical reactions during a process [180]. In plants, for example, mechanistic models have successfully predicted specific, observable traits such as plant height or processes like evapotransporation [175]. These models are then used as proxies for optimization targets like disease resistance and growth rates. In the early 2000s to 2010s, models were developed by [179] and others, which assumed fitness in any environment is best described by Entropy Production. While the MEP principle cannot be proved [181], it has been demonstrated as practical and applicable for both individual organisms and entire ecosystems [179,181,182]. More recently, Cannon et al. [180] used the MEP principle to model central metabolism in the fungus *Neurospora crassa* and used it to estimate kinetic rate constants normally obtained using painstaking experimentation [183]. This optimization objective used the MEP principle by choosing the parameters (inferred rate constants) that "can happen in the greatest number of ways" [180]. A hypothetical search space for such an optimization is depicted in Figure 7.

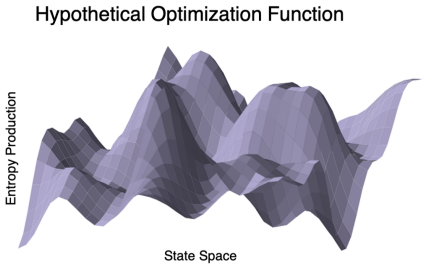

Figure 7. Biology presents many optimization problems. Shown here is a hypothetical search space that could represent anything from an organism's traits to experiment designs to visualization parameters. In these cases, the vertical axis could be fitness/MEP, experiment value, and network edge crossovers, respectively.

3.7.2. Dimensionality Reduction for Omics Analysis

Omics experiments suffer from the small-n-large-p problem, where a relatively small number of samples, n, each comprise a large number of colinear variables, p. Dimensionality reduction, reviewed by Sorzano et al. [184], amounts to the pursuit of a projection to lower-dimensional space that is *Efficient*, *Relevant*, and *Meaningful*. *Efficient* means the projection space has few dimensions and possibly weighs few of the n input variables (sparse or regularized), *Relevant* means that it maintains information related to the experiment, and *Meaningful* means that the projection is oriented in a useful or interesting position. While Principle Components Analysis (PCA) [185] finds the *Meaning* by projecting data to orthogonal axes ordered by their variance, Independent Component Analysis (ICA) [186] explicitly uses information theory to find *Meaning* differently. In ICA, the latent factors are assumed to be statistically independent and requires non-normality in the distributions of the independent latent factors. Reviewed in [187], this non-normality requirement is crucial to ICA and is optimized in different ways using different calculations of non-normality. One of those measures is *Negentropy*, J,

$$J(Y) = H(Z) - H(Y) \tag{18}$$

which is a measure of how different the differential (continuous) entropy H (Equation (4)) of a random variable Y is from that of a truly normal random variable Z with the same variance-covariance structure. The reason $H(Y)$ is subtracted from $H(Z)$ is given a fixed second moment (variance), the normal density has the greatest entropy. Further, minimum MI between projection axes is a good measure of maximum *Negetropy* [187]. An advantage of ICA is that latent factors need not be orthogonal, as shown in Figure 8. ICA is applied not only for dimensionality reduction, but also in signal processing [187]. A difference between the two is that ICA does not rank latent factors in order of importance, as PCA does. Lastly, ICA calculation is not deterministic, and can sometimes lead to unstable results.

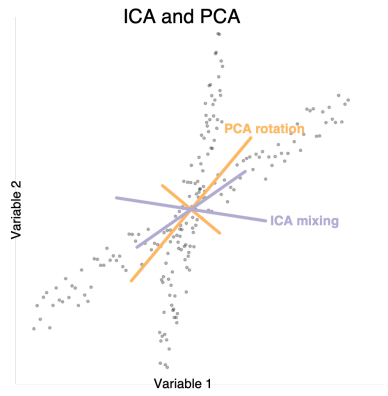

Figure 8. Independent Components Analysis (ICA) can compute latent factors in data like Principle Components Analysis (PCA). ICA is not limited to orthogonal mixtures like PCA, as shown in this simulated dataset. The ICA mixing vectors align to the data better than the PCA rotation vectors.

4. Discussion

In this article, we have performed a broad overview of applications of information theory in many key areas within the gamut of computational biology with focus on more recent developments (summarized in Tables 1 and 2). Information is intrinsically central to biology, most obviously because genetic information is stored in the DNA. Since the seminal work done by Shannon over seventy years ago, information theory with its foundations in statistical mechanics and communication theory, has made a tremendous impact to computational biology. Not only has it provided many ways of parsimoniously modeling and capturing non-linear associations between key components in biology (such as gene expression, DNA nucleotides, protein residues), it also helped represent dynamic biological systems as stochastic random processes. Its popularity also stems from the fact that information theory provides a strong alternative to many conventional statistical approaches often challenged with complex parameterization and computational intractability due to high dimensionality of input data. At the same time, although information theoretic applications have grown at a very fast pace and information theory based theoretical frameworks can be adapted to a wide range of problems, many challenges exist that should be considered prior to any analysis, as discussed in the context of biological applications in Andrews et al. [188] and for general data mining by Holzinger et al. [189]. The central concept in information theory is Shannon's entropy that is based on expected value of a probability distribution akin to statistical averaging. Because heterogeneity of traits and phenotypes is a rule in biology, rather than an exception, averages may not always work well in explaining or predicting behaviors. Many proposed information theory based test statistics should be used with caution as stationarity assumptions generally do not hold, sample sizes may be too small to support the law of large numbers and asymptotic properties are not well understood. Estimation of metrics such as interaction information relies on robust empirical estimations of multivariate distributions and joint probability mass functions with groups of random variables that are not straightforward at all, and more so, when data points are limited. Difficulties related to obtaining sufficient sample sizes and the computational burden associated with such estimations using high-dimensional and heterogeneous data often encountered in biology can result in bottlenecks in the application of information theory to systems biology [190]. Another set of issues to be tackled exist in the modeling front that include behavior of noise and robustness of models to imperfections and irregularities, a frequent occurrence in the experimental biological domain [188,191]. Assumptions of ideal conditions (infinite block lengths, additive white Gaussian noise, i.i.d distributions, etc.) underlying many classical information theoretic results may not always hold when the conditions are relaxed. Finally, some other open questions are how to select an appropriate entropy measure, its parameters and higher order metric(s) using the entropy measure to address a particular problem in biology and how to generally benchmark multiple such entropy measures [189]. More research is required addressing the above-mentioned challenges within specific applications in biological domains.

Table 1. Summary of key information theoretic methods for areas discussed in Sections 3.1–3.4.

Area	Information Theoretic Methods	How Information Theory Is Used
Reconstructing Gene Regulatory Networks	Relevance Networks [34]	Used MI larger than a given threshold to construct GRNs
	CLR [36]	Used MI to construct GRN, filters spurious edges by estimating its background distribution
	ARACNE and its extensions [37–40]	Used MI to construct GRN, filters edges using null distribution and DPI, higher order DPI to improve inference performance, adaptive binning strategy to estimate MI efficiently
	PIDC [27]	GRN constructed using PID to explore dependencies between triplets of genes in single-cell gene expression datasets
Alignment-free phylogeny	FFP [54]	Calculated JSD as pairwise distances between two genomes using normalized k-mer frequencies
	kWIP [55]	Constructed a sketch data structure using all k-mers from a genomic sequence, computed inner product between the two sketches weighted by their Shannon's entropy across the given dataset
Sequencing and Error Correction	Motahari et al. [66]	Used Renyi entropy of order 2 to find the minimum fragment length and coverage depth needed for the assembling reads to reconstruct the original DNA sequence with a given reliability
	Chen et al. [65]	Analyzed the information redundancy in dual-base degenerate sequencing by comparing entropy information content of multiple DPL (degenerate polymer length) arrays
	Anavy et al. [77]	Proposed encoding and decoding for composite DNA based storage and error correction through developing composite DNA alphabets and using KLD to select the best alphabet model
	Choi et al. [77]	Used eleven degenerate bases as encoding characters in addition to ACGT to increase information capacity limit and reduce the cost of DNA per unit data
Genome-wide disease-gene association mapping	Fan et al. [97], Yee et al. [98], Dong et al. [99]	Proposed test statistics based on MI and conditional entropy involving two SNPs and the disease phenotype from case-control studies
	Varadan et al. [103], Anastassiou [104], Curk at al. [105], Hu et al. [106,108]	Used synergy to analyze GXG statistical interactions
	Chanda et al. [24,114,115,118], Tritchler et al. [120]	Used multivariate information theoretic metrics and higher order models (e.g., KWII) to analyze statistical GXG and GxE interactions
	Moore et al. [124]	Used KWII to represent edges in epistasis networks
	Tahmasebi et al. [133]	Used entropy to formulate the capacity of recovering the causal subsequence
	Chanda et al. [116], Knights et al. [117], Yee et al. [131], Galas et al. [132]	Discussion and analysis of Information theoretic methods dealing with quantitative phenotypes and environmental variables
	Andrade et al. [128], Kang et al. [127]	Developed information theoretic test statistics for single-group or case-only studies.
	Tzeng et al. [92], Zhao et al. [93,94]	Developed entropy-based tests using haplotypes
	Cui et al. [95]	Developed entropy-based association test for gene-centric analysis considering variants within one gene as testing units
	Zhao et al. [130], Brunel et al. [102], Wu et al. [88]	Designed and discussed entropy based disease association test metrics for family-based studies

In the past few years, cutting-edge and high throughput technologies and experimental capabilities in biology have enabled rapid collection of unprecedented amounts of data ranging from millions of genomic sequences, images of physiological structures, high resolution microscopy images of cell morphologies to petabytes of health records. Therefore, it is worthwhile to briefly mention the role of information theory in the context of the recent mega-trends in data generation and analytics, particularly with respect to the newer paradigms of "Big Data" and "Deep Learning" that have emerged over the past two decades. The 3V's ("Volume", "Velocity" and "Variety") constitute three key properties of any "Big Data" in general that informally refers to the data deluge that generates large "Volumes" of data at high "Velocities" with a lot of "Variety" [192,193]. Within biology, genomics, through production of high-throughput sequencing data, continues to lead in terms of data growth and availability [192,194]. Information theoretic approaches, being at the heart of some of the popular pattern mining and statistical machine learning methods, stand to benefit from the increasing "Volume" and "Velocity" primarily through more reliable estimations of underlying empirical data distributions leading to improved estimation of metrics such as MI and KWII. However, ever increasing production of high-throughput biological data poses serious challenges to the conventional solutions for storing, processing and transmitting these data. Consequently, in addition to dimensionality reduction strategies for omics analysis, DNA based data storage [72] and data compression methods [195] for various biological data types have become active areas of research over the last decade. Because of the long-term stability and information density, using DNA as an archival medium is on track to become an appealing medium for handling next generation data "Volume". Multiple methods have been proposed in the last few years [72,77,78,85–87] that aim to harness the properties of DNA to mathematically model it as a storage channel and improve it's information capacity. In the data compression area, some of the early compression methods, such as GenCompress [196] used reference genomes to map short sequences and then used entropy coding algorithms to encode the addresses of the short sequences, their lengths and their probable substitutions. More recent methods like CoGI [197] and iDoComp [198] can be used for both reference-free as well as for reference-based genome compression and rely on applying advanced data structures and algorithms such as suffix arrays and rectangular partition coding [199] to reduce mapping size before using entropy encoding in the final compression step. Another promising direction to manage this data deluge can be found in the work of Yu et al. [200] and later advanced by Ishaq et al. [201]. Their research proposed information theory and hierarchical clustering based framework for similarity search in massive biological datasets based on characterizing a dataset's entropy and fractal dimensions, and enabled reduction in data volume by pruning redundant data while preserving the essential structures and patterns within the data. This not only attempts to address data "Volume", it also holds the potential to improve data "Veracity" (from 5V categorization of "Big data" [193]) that reduces data redundancy and noise and can lead to improved estimations of information theoretic as well as statistical metrics in downstream applications. Fueled by the ever-increasing growth in biological data and emerging techniques such as CRISPR based genome-editing, we are only beginning to explore the capacity of information science to advance research in biological data compression and DNA based storage, and we expect to see many more exciting developments in these areas.

Table 2. Summary of key information theoretic methods for areas discussed in Sections 3.5–3.7.

Area	Information Theoretic Methods	How Information Theory Is Used
Reconstruction and analysis of Metabolic Networks	CLR [36], ARACNE [37], PCLRC [164]	Uses MI to construct metabolic networks from metabolite concentration data, filters spurious edges by estimating its background distribution [36,164] or DPI [37]
	Marr. et al. [166,167]	Network analysis of metabolic networks using Shannon and word entropy to reveal regularization dynamics encoded in network topology
	Nykter et al. [168]	Studied network structure-dynamics relationships, using Kolmogorov complexity as a measure of distance between pairs of network structures and between their associated dynamic state trajectories
	Grimbs et al. [169]	Stoichiometric analysis to parameterize the metabolic states, assessed the effect of enzyme-kinetic parameters on the stability properties of a metabolic state using MI and Kolmogorov–Smirnov test
	Fuente et al. [11].	Studied properties of dissipative metabolic structures at different organizational levels using entropy
	De Martino et al. [170]	Introduced a generalization of FBA to single-cell level based on maximum entropy principle
	Saccenti et al. [163]	Investigated the associations and the interconnections among different metabolites by means of network modeling using maximum entropy ensemble null model
	Wagner et al. [172]	Proposed an information theoretic way to relate the nutrient information to the error in a cell's measurement of nutrient concentration in its environment
Protein interaction analysis	Wollenberg et al. [154], Tillier et al. [155], Dunn et al. [156], Kamisetty et al. [159], Morcos et al. [158]	Mutual Information combined with evolutionary information and refined with structural information to identify protein interactions
Optimization, Dimensionality Reduction	Cannon et al. [202,203], Thomas et al. [204]	Used MEP to simulate central metabolism in the fungus Neurospora crassa [180], tricarboxylic acid cycle model optimization in microbes [204].
	Hyv´arinen et al. [187], Comon et al. [186]	Used negentropy and minimization of MI to obtain the components in ICA

More recently, inspired by knowledge of neural information processing and functioning of the brain, and powered by availability of modern powerful GPUs, artificial intelligence and deep learning have made impressive advances in numerous applications ranging from computer vision, natural-language and speech processing to bioinformatics and computational biology. Keeping up with this trend, information theory, in many forms, has now become ubiquitous within many state-of-the-art algorithms in machine learning and deep learning algorithms. It has helped advance theoretical developments of optimization and training in deep learning, that drives many emerging applications in biological data science such as genomic predictions and imaging genomics exploring relationships between genotypes, phenotypes and clinical outcomes [205,206]. For example, cross entropy has become a standard for comparing two probability distributions and is a popular loss function for deep neural networks, in both binomial and multinomial classification scenarios. Information theory is also continuing to play a central role in on-going research investigating information bottleneck and related principles in the analysis and design of representation learning and optimization algorithms for training deep neural networks, for example, as in [207].

Lastly, with recent advancements in quantum computations, Shannon's classical information theory is paving the way for using quantum information theory with physics to further the understanding of biological systems. Active research is being pursued in many uses of quantum information theory, such as developing quantum biological channel models suitable for the study of quantum information transfer from DNA to proteins [208], quantum-mechanical modeling of spontaneous, induced, and adaptive mutations and their role in cancerous tumor developments [209] and using both classical and quantum error-correction coding in genetics and evolution [210]. We conjecture that information theory has a key role to play in many theoretical and application developments in these areas in the near future.

Author Contributions: P.C. conceived the idea and organization of the review article. P.C. contributed to Section 2, Sections 3.1, 3.2, 3.4 and 4; E.C. contributed to Sections 3.3 and 3.4; S.S. and R.W. contributed to Section 3.5; J.H. contributed to Section 3.6 and J.V.H. contributed to Section 3.7. The author names are ordered alphabetically in order of the last names. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Authors are grateful for the support and feedback from Siva Kumpatla, Kevin Hayes and Jochen Scheel from Data Science and Informatics, Corteva Agriscience™ leadership team.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
3. Tsimring, L.S. Noise in biology. *Rep. Prog. Phys.* **2014**, *77*, 026601. [CrossRef] [PubMed]
4. Mousavian, Z.; Kavousi, K.; Masoudi-Nejad, A. *Information Theory in Systems Biology. Part I: Gene Regulatory And Metabolic Networks*; Seminars in Cell & Developmental Biology; Elsevier: Amsterdam, The Netherlands; 2016; Volume 51, pp. 3–13.
5. Mousavian, Z.; Díaz, J.; Masoudi-Nejad, A. *Information Theory in Systems Biology. Part II: Protein–Protein Interaction and Signaling Networks*; Seminars in Cell & Developmental Biology; Elsevier: Amsterdam, The Netherlands, 2016; Volume 51, pp. 14–23.
6. Vinga, S. Information theory applications for biological sequence analysis. *Brief. Bioinform.* **2014**, *15*, 376–389. [CrossRef] [PubMed]
7. Waltermann, C.; Klipp, E. Information theory based approaches to cellular signaling. *Biochim. Biophys. Acta* **2011**, *1810*, 924–932. [CrossRef] [PubMed]
8. Chen, S.; Mar, J.C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinform.* **2018**, *19*, 232. [CrossRef] [PubMed]
9. Zielezinski, A.; Girgis, H.Z.; Bernard, G.; Leimeister, C.A.; Tang, K.; Dencker, T.; Lau, A.K.; Röhling, S.; Choi, J.J.; Waterman, M.S.; et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **2019**, *20*, 144. [CrossRef]

10. Little, D.Y.; Chen, L. Identification of Coevolving Residues and Coevolution Potentials Emphasizing Structure, Bond Formation and Catalytic Coordination in Protein Evolution. *PLoS ONE* **2009**, *4*, e4762. [CrossRef]
11. Martínez de la Fuente, I. Quantitative analysis of cellular metabolic dissipative, self-organized structures. *Int. J. Mol. Sci.* **2010**, *11*, 3540–3599. [CrossRef]
12. Schneider, T.D. A brief review of molecular information theory. *Nano Commun. Netw.* **2010**, *1*, 173–180. [CrossRef]
13. Chen, H.D.; Chang, C.H.; Hsieh, L.C.; Lee, H.C. Divergence and Shannon information in genomes. *Phys. Rev. Lett.* **2005**, *94*, 178103. [CrossRef] [PubMed]
14. Chang, C.H.; Hsieh, L.C.; Chen, T.Y.; Chen, H.D.; Luo, L.; Lee, H.C. Shannon information in complete genomes. *J. Bioinform. Comput. Biol.* **2005**, *3*, 587–608. [CrossRef] [PubMed]
15. Machado, J.T.; Costa, A.C.; Quelhas, M.D. Shannon, Rényie and Tsallis entropy analysis of DNA using phase plane. *Nonlinear Anal. Real World Appl.* **2011**, *12*, 3135–3144. [CrossRef]
16. Athanasopoulou, L.; Athanasopoulos, S.; Karamanos, K.; Almirantis, Y. Scaling properties and fractality in the distribution of coding segments in eukaryotic genomes revealed through a block entropy approach. *Phys. Rev. E* **2010**, *82*, 051917. [CrossRef] [PubMed]
17. Vinga, S. Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification. *Adv. Comput. Methods Biocomput. Bioimaging* **2007**, *71*, 107.
18. Vinga, S.; Almeida, J. Alignment-free sequence comparison—A review. *Bioinformatics* **2003**, *19*, 513–523. [CrossRef]
19. Ladbury, J.E.; Arold, S.T. Noise in cellular signaling pathways: Causes and effects. *Trends Biochem. Sci.* **2012**, *37*, 173–178. [CrossRef]
20. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin, Germany, 2006.
21. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall PTR: London, UK, 1994.
22. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
23. Jakulin, A. Machine Learning Based on Attribute Interactions. Ph.D. Thesis, Univerza v Ljubljani, Ljubljana, Republika Slovenija, 2005.
24. Chanda, P.; Zhang, A.; Brazeau, D.; Sucheston, L.; Freudenheim, J.L.; Ambrosone, C.; Ramanathan, M. Information-theoretic metrics for visualizing gene-environment interactions. *Am. J. Hum. Genet.* **2007**, *81*, 939–963. [CrossRef]
25. Te Sun, H. Multiple mutual informations and multiple interactions in frequency data. *Inf. Control* **1980**, *46*, 26–45.
26. Trapnell, C.; Cacchiarelli, D.; Grimsby, J.; Pokharel, P.; Li, S.; Morse, M.; Lennon, N.J.; Livak, K.J.; Mikkelsen, T.S.; Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **2014**, *32*, 381. [CrossRef]
27. Chan, T.E.; Stumpf, M.P.; Babtie, A.C. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **2017**, *5*, 251–267. [CrossRef] [PubMed]
28. Moris, N.; Pina, C.; Arias, A.M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **2016**, *17*, 693. [CrossRef] [PubMed]
29. Zambelli, F.; Mastropasqua, F.; Picardi, E.; D'Erchia, A.M.; Pesole, G.; Pavesi, G. RNentropy: An entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Res.* **2018**, *46*, e46–e46. [CrossRef] [PubMed]
30. Qiu, X.; Rahimzamani, A.; Wang, L.; Mao, Q.; Durham, T.; McFaline-Figueroa, J.L.; Saunders, L.; Trapnell, C.; Kannan, S. Towards inferring causal gene regulatory networks from single cell expression measurements. *BioRxiv* **2018**. [CrossRef]
31. Meyer, P.; Kontos, K.; Lafitte, F.; Bontempi, G. EURASIP J. Bioinf. *Syst. Biol.* **2007**, *79879*.
32. Chaitankar, V.; Ghosh, P.; Perkins, E.J.; Gong, P.; Deng, Y.; Zhang, C. A novel gene network inference algorithm using predictive minimum description length approach. *BMC Syst. Biol.* **2010**, *4*, S7. [CrossRef]
33. Zhang, X.; Zhao, X.M.; He, K.; Lu, L.; Cao, Y.; Liu, J.; Hao, J.K.; Liu, Z.P.; Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **2012**, *28*, 98–104. [CrossRef]
34. Butte, A.J.; Kohane, I.S. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*; World Scientific: Singapore, 1999; pp. 418–429.

35. Butte, A.J.; Tamayo, P.; Slonim, D.; Golub, T.R.; Kohane, I.S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12182–12186. [CrossRef] [PubMed]
36. Faith, J.J.; Hayete, B.; Thaden, J.T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J.J.; Gardner, T.S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **2007**, *5*, e8. [CrossRef] [PubMed]
37. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context*; BMC Bioinformatics; Springer: Berlin, Germany, 2006; Volume 7, p. S7.
38. Zoppoli, P.; Morganella, S.; Ceccarelli, M. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinform.* **2010**, *11*, 154. [CrossRef] [PubMed]
39. Jang, I.S.; Margolin, A.; Califano, A. hARACNe: Improving the accuracy of regulatory model reverse engineering via higher-order data processing inequality tests. *Interface Focus* **2013**, *3*, 20130011. [CrossRef] [PubMed]
40. Lachmann, A.; Giorgi, F.M.; Lopez, G.; Califano, A. ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **2016**, *32*, 2233–2235. [CrossRef] [PubMed]
41. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
42. Vân Anh Huynh-Thu, A.I.; Wehenkel, L.; Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **2010**, *5*, e12776.
43. Matsumoto, H.; Kiryu, H.; Furusawa, C.; Ko, M.S.; Ko, S.B.; Gouda, N.; Hayashi, T.; Nikaido, I. SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **2017**, *33*, 2314–2321. [CrossRef] [PubMed]
44. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Aderhold, A.; Bonneau, R.; Chen, Y.; et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796. [CrossRef] [PubMed]
45. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186. [CrossRef]
46. Bonham-Carter, O.; Steele, J.; Bastola, D. Alignment-free genetic sequence comparisons: A review of recent approaches by word analysis. *Brief. Bioinform.* **2014**, *15*, 890–905. [CrossRef]
47. Wang, Y.; Liu, L.; Chen, L.; Chen, T.; Sun, F. Comparison of metatranscriptomic samples based on k-tuple frequencies. *PLoS ONE* **2014**, *9*, e84348. [CrossRef]
48. Wen, J.; Zhang, Y. A 2D graphical representation of protein sequence and its numerical characterization. *Chem. Phys. Lett.* **2009**, *476*, 281–286. [CrossRef]
49. Randić, M.; Zupan, J.; Balaban, A.T. Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **2004**, *397*, 247–252. [CrossRef]
50. Jeffrey, H.J. Chaos game representation of gene structure. *Nucleic Acids Res.* **1990**, *18*, 2163–2170. [CrossRef] [PubMed]
51. Almeida, J.S. Sequence analysis by iterated maps, a review. *Brief. Bioinform.* **2014**, *15*, 369–375. [CrossRef] [PubMed]
52. Leimeister, C.A.; Boden, M.; Horwege, S.; Lindner, S.; Morgenstern, B. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics* **2014**, *30*, 1991–1999. [CrossRef] [PubMed]
53. Morgenstern, B.; Zhu, B.; Horwege, S.; Leimeister, C.A. Estimating evolutionary distances between genomic sequences from spaced-word matches. *Algorithms Mol. Biol.* **2015**, *10*, 5. [CrossRef] [PubMed]
54. Sims, G.E.; Jun, S.R.; Wu, G.A.; Kim, S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 2677–2682. [CrossRef]
55. Murray, K.D.; Webers, C.; Ong, C.S.; Borevitz, J.; Warthmann, N. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS Comput. Biol.* **2017**, *13*, e1005727. [CrossRef]
56. Zhang, Q.; Pell, J.; Canino-Koning, R.; Howe, A.C.; Brown, C.T. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PLoS ONE* **2014**, *9*, e101271. [CrossRef]
57. Cormode, G.; Muthukrishnan, S. An improved data stream summary: The count-min sketch and its applications. *J. Algorithms* **2005**, *55*, 58–75. [CrossRef]

58. Drouin, A.; Giguère, S.; Déraspe, M.; Marchand, M.; Tyers, M.; Loo, V.G.; Bourgault, A.M.; Laviolette, F.; Corbeil, J. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genom.* **2016**, *17*, 1–15. [CrossRef]
59. Glouzon, J.P.S.; Perreault, J.P.; Wang, S. The super-n-motifs model: A novel alignment-free approach for representing and comparing RNA secondary structures. *Bioinformatics* **2017**, *33*, 1169–1178. [CrossRef] [PubMed]
60. Sarmashghi, S.; Bohmann, K.; Gilbert, M.T.P.; Bafna, V.; Mirarab, S. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biol.* **2019**, *20*, 1–20. [CrossRef] [PubMed]
61. Rhoads, A.; Au, K.F. PacBio sequencing and its applications. *Genom. Proteom. Bioinform* **2015**, *13*, 278–289. [CrossRef] [PubMed]
62. Laver, T.; Harrison, J.; Oâ€™neill, P.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the performance of the oxford nanopore technologies minion. *Biomol Detect. Quantif.* **2015**, *3*, 1–8. [CrossRef] [PubMed]
63. Bansal, V.; Boucher, C. *Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? iScience* **2019**, *18*, 37. [CrossRef]
64. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333. [CrossRef] [PubMed]
65. Chen, Z.; Zhou, W.; Qiao, S.; Kang, L.; Duan, H.; Xie, X.S.; Huang, Y. Highly accurate fluorogenic DNA sequencing with information theory–based error correction. *Nat. Biotechnol.* **2017**, *35*, 1170. [CrossRef]
66. Motahari, A.S.; Bresler, G.; David, N. Information theory of DNA shotgun sequencing. *IEEE Trans. Inf. Theory* **2013**, *59*, 6273–6289. [CrossRef]
67. Vinga, S.; Almeida, J.S. Rényi continuous entropy of DNA sequences. *J. Theor. Biol.* **2004**, *231*, 377–388. [CrossRef]
68. Shomorony, I.; Courtade, T.; Tse, D. Do read errors matter for genome assembly? In Proceedings of the 2015 IEEE International Symposium on Information Theory (ISIT), Hong Kong, China, 14–19 June 2015; pp. 919–923.
69. Bresler, G.; Bresler, M.; Tse, D. *Optimal Assembly for High Throughput Shotgun Sequencing*; BMC Bioinformatics; Springer: Berlin, Germany, 2013, Volume 14, p. S18.
70. Ganguly, S.; Mossel, E.; Rácz, M.Z. Sequence assembly from corrupted shotgun reads. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 265–269.
71. Gabrys, R.; Milenkovic, O. Unique reconstruction of coded sequences from multiset substring spectra. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 2540–2544.
72. Shomorony, I.; Heckel, R. DNA-Based Storage: Models and Fundamental Limits. *arXiv* **2020**, arXiv:2001.06311.
73. Marcovich, S.; Yaakobi, E. Reconstruction of Strings from their Substrings Spectrum. *arXiv* **2019**, arXiv:1912.11108.
74. Si, H.; Vikalo, H.; Vishwanath, S. Information-theoretic analysis of haplotype assembly. *IEEE Trans. Inf. Theory* **2017**, *63*, 3468–3479. [CrossRef]
75. Sims, P.A.; Greenleaf, W.J.; Duan, H.; Xie, X.S. Fluorogenic DNA sequencing in PDMS microreactors. *Nat. Methods* **2011**, *8*, 575. [CrossRef]
76. Mitchell, K.; Brito, J.J.; Mandric, I.; Wu, Q.; Knyazev, S.; Chang, S.; Martin, L.S.; Karlsberg, A.; Gerasimov, E.; Littman, R.; et al. Benchmarking of computational error-correction methods for next-generation sequencing data. *Genome Biol.* **2020**, *21*, 1–13. [CrossRef]
77. Anavy, L.; Vaknin, I.; Atar, O.; Amit, R.; Yakhini, Z. Improved DNA based storage capacity and fidelity using composite DNA letters. *bioRxiv* **2018**. [CrossRef]
78. Choi, Y.; Ryu, T.; Lee, A.; Choi, H.; Lee, H.; Park, J.; Song, S.H.; Kim, S.; Kim, H.; Park, W.; et al. Addition of degenerate bases to DNA-based data storage for increased information capacity. *bioRxiv* **2018**. [CrossRef]
79. Reed, I.S.; Solomon, G. Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.* **1960**, *8*, 300–304. [CrossRef]
80. Fu, S.; Wang, A.; Au, K.F. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* **2019**, *20*, 26. [CrossRef] [PubMed]
81. Amarasinghe, S.L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M.E.; Gouil, Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **2020**, *21*, 1–16. [CrossRef]

82. Mantere, T.; Kersten, S.; Hoischen, A. Long-read sequencing emerging in medical genetics. *Front. Genet.* **2019**, *10*, 426. [CrossRef]
83. Nakano, K.; Shiroma, A.; Shimoji, M.; Tamotsu, H.; Ashimine, N.; Ohki, S.; Shinzato, M.; Minami, M.; Nakanishi, T.; Teruya, K.; et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum. Cell* **2017**, *30*, 149–161. [CrossRef] [PubMed]
84. Boldogkői, Z.; Moldován, N.; Balázs, Z.; Snyder, M.; Tombácz, D. Long-read sequencing—A powerful tool in viral transcriptome research. *Trends Microbiol.* **2019**, *27*, 578–592. [CrossRef]
85. Heckel, R.; Shomorony, I.; Ramchandran, K.; David, N. Fundamental limits of DNA storage systems. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 3130–3134.
86. Meiser, L.C.; Antkowiak, P.L.; Koch, J.; Chen, W.D.; Kohll, A.X.; Stark, W.J.; Heckel, R.; Grass, R.N. Reading and writing digital data in DNA. *Nat. Protoc.* **2020**, *15*, 86–101. [CrossRef] [PubMed]
87. Lopez, R.; Chen, Y.J.; Ang, S.D.; Yekhanin, S.; Makarychev, K.; Racz, M.Z.; Seelig, G.; Strauss, K.; Ceze, L. DNA assembly for nanopore data storage readout. *Nat. Commun.* **2019**, *10*, 1–9. [CrossRef] [PubMed]
88. Wu, C.; Li, S.; Cui, Y. Genetic association studies: an information content perspective. *Curr. Genom.* **2012**, *13*, 566–573. [CrossRef]
89. Kang, G.; Zuo, Y. Entropy-based joint analysis for two-stage genome-wide association studies. *J. Hum. Genet.* **2007**, *52*, 747–756. [CrossRef]
90. Ruiz-Marín, M.; Matilla-García, M.; Cordoba, J.A.G.; Susillo-González, J.L.; Romo-Astorga, A.; González-Pérez, A.; Ruiz, A.; Gayán, J. An entropy test for single-locus genetic association analysis. *BMC Genet.* **2010**, *11*, 19. [CrossRef]
91. Li, P.; Guo, M.; Wang, C.; Liu, X.; Zou, Q. An overview of SNP interactions in genome-wide association studies. *Brief. Funct. Genom.* **2015**, *14*, 143–155. [CrossRef]
92. Tzeng, J.Y.; Devlin, B.; Wasserman, L.; Roeder, K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.* **2003**, *72*, 891–902. [CrossRef]
93. Zhao, J.; Boerwinkle, E.; Xiong, M. An entropy-based statistic for genomewide association studies. *Am. J. Hum. Genet.* **2005**, *77*, 27–40. [CrossRef]
94. Zhao, J.; Jin, L.; Xiong, M. Nonlinear tests for genomewide association studies. *Genetics* **2006**, *174*, 1529–1538. [CrossRef] [PubMed]
95. Cui, Y.; Kang, G.; Sun, K.; Qian, M.; Romero, R.; Fu, W. Gene-centric genomewide association study via entropy. *Genetics* **2008**, *179*, 637–650. [CrossRef] [PubMed]
96. Cordell, H.J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **2002**, *11*, 2463–2468. [CrossRef] [PubMed]
97. Fan, R.; Zhong, M.; Wang, S.; Zhang, Y.; Andrew, A.; Karagas, M.; Chen, H.; Amos, C.; Xiong, M.; Moore, J. Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genet. Epidemiol.* **2011**, *35*, 706–721. [CrossRef]
98. Yee, J.; Kwon, M.S.; Park, T.; Park, M. A modified entropy-based approach for identifying gene-gene interactions in case-control study. *PLoS ONE* **2013**, *8*, e69321. [CrossRef]
99. Dong, C.; Chu, X.; Wang, Y.; Wang, Y.; Jin, L.; Shi, T.; Huang, W.; Li, Y. Exploration of gene–gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet.* **2008**, *16*, 229–235. [CrossRef]
100. Ferrario, P.G.; König, I.R. Transferring entropy to the realm of GxG interactions. *Brief. Bioinform.* **2018**, *19*, 136–147. [CrossRef]
101. Taylor, M.B.; Ehrenreich, I.M. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.* **2015**, *31*, 34–40. [CrossRef]
102. Brunel, H.; Gallardo-Chacón, J.J.; Buil, A.; Vallverdú, M.; Soria, J.M.; Caminal, P.; Perera, A. MISS: A non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics* **2010**, *26*, 1811–1818. [CrossRef]
103. Varadan, V.; Miller III, D.M.; Anastassiou, D. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* **2006**, *22*, e497–e506. [CrossRef]
104. Anastassiou, D. Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.* **2007**, *3*, 83. [CrossRef] [PubMed]
105. Curk, T.; Rot, G.; Zupan, B. SNPsyn: detection and exploration of SNP–SNP interactions. *Nucleic Acids Res.* **2011**, *39*, W444–W449. [CrossRef] [PubMed]

106. Hu, T.; Sinnott-Armstrong, N.A.; Kiralis, J.W.; Andrew, A.S.; Karagas, M.R.; Moore, J.H. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinform.* **2011**, *12*, 364. [CrossRef] [PubMed]
107. Hu, T.; Chen, Y.; Kiralis, J.W.; Collins, R.L.; Wejse, C.; Sirugo, G.; Williams, S.M.; Moore, J.H. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *J. Am. Med. Inf. Assoc.* **2013**, *20*, 630–636. [CrossRef]
108. Hu, T.; Chen, Y.; Kiralis, J.W.; Moore, J.H. Vi SEN: Methodology and Software for Visualization of Statistical Epistasis Networks. *Genet. Epidemiol.* **2013**, *37*, 283–285. [CrossRef]
109. Lee, W.; Sjölander, A.; Pawitan, Y. A critical look at entropy-based gene-gene interaction measures. *Genet. Epidemiol.* **2016**, *40*, 416–424. [CrossRef]
110. Shang, J.; Zhang, J.; Sun, Y.; Zhang, Y. EpiMiner: a three-stage co-information based method for detecting and visualizing epistatic interactions. *Digit. Signal Process.* **2014**, *24*, 1–13. [CrossRef]
111. Mielniczuk, J.; Rdzanowski, M. Use of information measures and their approximations to detect predictive gene-gene interaction. *Entropy* **2017**, *19*, 23. [CrossRef]
112. Chen, L.; Yu, G.; Langefeld, C.D.; Miller, D.J.; Guy, R.T.; Raghuram, J.; Yuan, X.; Herrington, D.M.; Wang, Y. Comparative analysis of methods for detecting interacting loci. *BMC Genom.* **2011**, *12*, 344. [CrossRef]
113. Chen, G.; Yuan, A.; Cai, T.; Li, C.M.; Bentley, A.R.; Zhou, J.; N. Shriner, D.; A. Adeyemo, A.; N. Rotimi, C. Measuring gene–gene interaction using Kullback–Leibler divergence. *Ann. Hum. Genet.* **2019**, *83*, 405–417. [CrossRef]
114. Chanda, P.; Sucheston, L.; Zhang, A.; Brazeau, D.; Freudenheim, J.L.; Ambrosone, C.; Ramanathan, M. AMBIENCE: A novel approach and efficient algorithm for identifying informative genetic and environmental associations with complex phenotypes. *Genetics* **2008**, *180*, 1191–1210. [CrossRef] [PubMed]
115. Chanda, P.; Sucheston, L.; Zhang, A.; Ramanathan, M. The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors. *Eur. J. Hum. Genet.* **2009**, *17*, 1274–1286. [CrossRef]
116. Chanda, P.; Sucheston, L.; Liu, S.; Zhang, A.; Ramanathan, M. Information-theoretic gene-gene and gene-environment interaction analysis of quantitative traits. *BMC Genom.* **2009**, *10*, 509. [CrossRef] [PubMed]
117. Knights, J.; Yang, J.; Chanda, P.; Zhang, A.; Ramanathan, M. SYMPHONY, an information-theoretic method for gene–gene and gene–environment interaction analysis of disease syndromes. *Heredity* **2013**, *110*, 548–559. [CrossRef] [PubMed]
118. Chanda, P.; Zhang, A.; Ramanathan, M. Modeling of environmental and genetic interactions with AMBROSIA, an information-theoretic model synthesis method. *Heredity* **2011**, *107*, 320–327. [CrossRef]
119. Knights, J.; Ramanathan, M. An information theory analysis of gene-environmental interactions in count/rate data. *Hum. Hered.* **2012**, *73*, 123–138. [CrossRef]
120. Tritchler, D.L.; Sucheston, L.; Chanda, P.; Ramanathan, M. Information metrics in genetic epidemiology. *Stat. Appl. Genet. Mol. Biol.* **2011**, *10*. [CrossRef]
121. Sucheston, L.; Chanda, P.; Zhang, A.; Tritchler, D.; Ramanathan, M. Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity. *BMC Genom.* **2010**, *11*, 487. [CrossRef]
122. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **2001**, *69*, 138–147. [CrossRef]
123. Culverhouse, R. The use of the restricted partition method with case-control data. *Hum. Hered.* **2007**, *63*, 93–100. [CrossRef]
124. Moore, J.H.; Hu, T. Epistasis analysis using information theory. In *Epistasis*; Springer: Berlin, Germany, 2015; pp. 257–268.
125. Barabási, A.L.; Bonabeau, E. Scale-free networks. *Sci. Am.* **2003**, *288*, 60–69. [CrossRef] [PubMed]
126. Piegorsch, W.W.; Weinberg, C.R.; Taylor, J.A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **1994**, *13*, 153–162. [CrossRef] [PubMed]
127. Kang, G.; Yue, W.; Zhang, J.; Cui, Y.; Zuo, Y.; Zhang, D. An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol.* **2008**, *250*, 362–374. [CrossRef] [PubMed]

128. De Andrade, M.; Wang, X. Entropy based genetic association tests and gene-gene interaction tests. *Stat. Appl. Genet. Mol. Biol.* **2011**, *10*. [CrossRef]
129. Spielman, R.S.; McGinnis, R.E.; Ewens, W.J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **1993**, *52*, 506.
130. Zhao, J.; Boerwinkle, E.; Xiong, M. An entropy-based genome-wide transmission/disequilibrium test. *Hum. Genet.* **2007**, *121*, 357–367. [CrossRef]
131. Yee, J.; Kwon, M.S.; Jin, S.; Park, T.; Park, M. Detecting Genetic Interactions for Quantitative Traits Using-Spacing Entropy Measure. *BioMed. Res. Int.* **2015**, *2015*, 523641. [CrossRef]
132. Galas, D.J.; Kunert-Graf, J.M.; Uechi, L.; Sakhanenko, N.A. Towards an information theory of quantitative genetics. *bioRxiv* **2019**, 811950.
133. Tahmasebi, B.; Maddah-Ali, M.A.; Motahari, A.S. Genome-wide association studies: Information theoretic limits of reliable learning. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 2231–2235.
134. Tahmasebi, B.; Maddah-Ali, M.A.; Motahari, S.A. Information Theory of Mixed Population Genome-Wide Association Studies. In Proceedings of the 2018 IEEE Information Theory Workshop (ITW), Guangzhou, China, 25–29 November 2018; pp. 1–5.
135. Jiang, D.; Wang, M. Recent developments in statistical methods for GWAS and high-throughput sequencing association studies of complex traits. *Biostat. Epidemiol.* **2018**, *2*, 132–159. [CrossRef]
136. Hayes, B. Overview of statistical methods for genome-wide association studies (GWAS). In *Genome-Wide Association Studies and Genomic Prediction*; Springer: Berlin, Germany, 2013; pp. 149–169.
137. Kubkowski, M.; Mielniczuk, J. Asymptotic distributions of empirical Interaction Information. *Methodol. Comput. Appl. Probab.* **2020**, 1–25. [CrossRef]
138. Goeman, J.J.; Solari, A. Multiple hypothesis testing in genomics. *Stat. Med.* **2014**, *33*, 1946–1978. [CrossRef] [PubMed]
139. Chanda, P.; Zhang, A.; Ramanathan, M. Algorithms for Efficient Mining of Statistically Significant Attribute Association Information. *arXiv* **2012**, arXiv:1208.3812.
140. Wang, Y.; Liu, G.; Feng, M.; Wong, L. An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* **2011**, *27*, 2936–2943. [CrossRef]
141. Sevimoglu, T.; Arga, K.Y. The role of protein interaction networks in systems biomedicine. *Comput. Struct. Biotechnol. J.* **2014**, *11*, 22–27. [CrossRef]
142. De Las Rivas, J.; Fontanillo, C. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* **2010**, *6*, 523641 . [CrossRef]
143. Braun, P.; Gingras, A.C. History of protein–protein interactions: From egg-white to complex networks. *Proteomics* **2012**, *12*, 1478–1498. [CrossRef] [PubMed]
144. Droit, A.; Poirier, G.G.; Hunter, J.M. Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function. *J. Mol. Endocrinol.* **2005**, *34*, 263–280. [CrossRef] [PubMed]
145. Shoemaker, B.A.; Panchenko, A.R. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.* **2007**, *3*, e42. [CrossRef] [PubMed]
146. Xing, S.; Wallmeroth, N.; Berendzen, K.W.; Grefen, C. Techniques for the analysis of protein-protein interactions in vivo. *Plant Phys.* **2016**, *171*, 727–758. [CrossRef]
147. Jansen, R.; Yu, H.; Greenbaum, D.; Kluger, Y.; Krogan, N.J.; Chung, S.; Emili, A.; Snyder, M.; Greenblatt, J.F.; Gerstein, M. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science* **2003**, *302*, 449–453. [CrossRef]
148. Pržulj, N. Protein-protein interactions: Making sense of networks via graph-theoretic modeling. *BioEssays* **2010**, *33*, 115–123. [CrossRef] [PubMed]
149. Fryxell, K.J. The coevolution of gene family trees. *Trends Genet.* **1996**, *12*, 364–369. [CrossRef]
150. Pazos, F.; Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng. Des. Sel.* **2001**, *14*, 609–614. [CrossRef] [PubMed]
151. Pazos, F.; Ranea, J.A.; Juan, D.; Sternberg, M.J. Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome. *J. Mol. Biol.* **2005**, *352*, 1002–1015. [CrossRef]
152. Fraser, H.B.; Hirsh, A.E.; Wall, D.P.; Eisen, M.B. Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9033–9038. [CrossRef]

153. Giraud, B.G.; Lapedes, A.; Liu, L.C. Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E* **1998**, *58*, 6312–6322. [CrossRef]
154. Wollenberg, K.R.; Atchley, W.R. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3288–3291. [CrossRef]
155. Tillier, E.R.; Lui, T.W. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* **2003**, *19*, 750–755. [CrossRef]
156. Dunn, S.; Wahl, L.; Gloor, G. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **2007**, *24*, 333–340. [CrossRef]
157. Szurmant, H.; Weigt, M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr. Opin. Struct. Biol.* **2018**, *50*, 26–32. [CrossRef]
158. Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D.S.; Sander, C.; Zecchina, R.; Onuchic, J.N.; Hwa, T.; Weigt, M.; et al.. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **2011**, *108*. [CrossRef] [PubMed]
159. Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15674–15679. [CrossRef] [PubMed]
160. Cong, Q.; Anishchenko, I.; Ovchinnikov, S.; Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **2019**, *365*, 185–189. [CrossRef] [PubMed]
161. Rosato, A.; Tenori, L.; Cascante, M.; De Atauri Carulla, P.R.; Martins dos Santos, V.A.P.; Saccenti, E. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics* **2018**, *14*, 37. [CrossRef] [PubMed]
162. Cakır, T.; Hendriks, M.M.; Westerhuis, J.A.; Smilde, A.K. Metabolic network discovery through reverse engineering of metabolome data. *Metabolomics* **2009**, *5*, 318–329. [CrossRef] [PubMed]
163. Saccenti, E.; Menichetti, G.; Ghini, V.; Remondini, D.; Tenori, L.; Luchinat, C. Entropy-based network representation of the individual metabolic phenotype. *J. Proteome Res.* **2016**, *15*, 3298–3307. [CrossRef] [PubMed]
164. Saccenti, E.; Suarez-Diez, M.; Luchinat, C.; Santucci, C.; Tenori, L. Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk. *J. Proteome Res.* **2015**, *14*, 1101–1111. [CrossRef]
165. Everett, J.R.; Holmes, E.; Veselkov, K.A.; Lindon, J.C.; Nicholson, J.K. A unified conceptual framework for metabolic phenotyping in diagnosis and prognosis. *Trends Pharmacol. Sci.* **2019**, *40*, 763–773. [CrossRef]
166. Marr, C.; Hütt, M.T. Topology regulates pattern formation capacity of binary cellular automata on graphs. *Phys. A Stat. Mech. Appl.* **2005**, *354*, 641–662. [CrossRef]
167. Marr, C.; Müller-Linow, M.; Hütt, M.T. Regularizing capacity of metabolic networks. *Phys. Rev. E* **2007**, *75*, 041917. [CrossRef]
168. Nykter, M.; Price, N.D.; Larjo, A.; Aho, T.; Kauffman, S.A.; Yli-Harja, O.; Shmulevich, I. Critical networks exhibit maximal information diversity in structure-dynamics relationships. *Phys. Rev. Lett.* **2008**, *100*, 058702. [CrossRef] [PubMed]
169. Grimbs, S.; Selbig, J.; Bulik, S.; Holzhütter, H.G.; Steuer, R. The stability and robustness of metabolic states: identifying stabilizing sites in metabolic networks. *Mol. Syst. Biol.* **2007**, *3*, 146. [CrossRef] [PubMed]
170. De Martino, D.; Mc Andersson, A.; Bergmiller, T.; Guet, C.C.; Tkačik, G. Statistical mechanics for metabolic networks during steady state growth. *Nat. Commun.* **2018**, *9*, 2988. [CrossRef] [PubMed]
171. Shore, J.; Johnson, R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory* **1980**, *26*, 26–37. [CrossRef]
172. Wagner, A. From bit to it: How a complex metabolic network transforms information into living matter. *BMC Syst. Biol.* **2007**, *1*, 33. [CrossRef]
173. Heirendt, L.; Arreckx, S.; Pfau, T.; Mendoza, S.N.; Richelle, A.; Heinken, A.; Haraldsdóttir, H.S.; Wachowiak, J.; Keating, S.M.; Vlasov, V.; et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **2019**, *14*, 639–702. [CrossRef]
174. Covert, M.W.; Schilling, C.H.; Famili, I.; Edwards, J.S.; Goryanin, I.I.; Selkov, E.; Palsson, B.O. Metabolic modeling of microbial strains in silico. *Trends Biochem. Sci.* **2001**, *26*, 179–186. [CrossRef]
175. Hammer, G.; Cooper, M.; Tardieu, F.; Welch, S.; Walsh, B.; van Eeuwijk, F.; Chapman, S.; Podlich, D. Models for navigating biological complexity in breeding improved crop plants. *Trends Plant Sci.* **2006**, *11*, 587–593. [CrossRef]

176. Gomes de Oliveira Dal'Molin, C.; Quek, L.E.; Saa, P.A.; Nielsen, L.K. A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Front. Plant Sci.* **2015**, *6*, 1–12. [CrossRef]
177. Sen, P.; Orešič, M. Metabolic modeling of human gut microbiota on a genome scale: An overview. *Metabolites* **2019**, *9*, 22. [CrossRef]
178. Chen, Y.; Li, G.; Nielsen, J. Genome-Scale Metabolic Modeling from Yeast to Human Cell Models of Complex Diseases: Latest Advances and Challenges. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2019; Volume 2049, pp. 329–345. [CrossRef]
179. Dewar, R.C. Maximum entropy production and plant optimization theories. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 1429–1435. [CrossRef] [PubMed]
180. Cannon, W.; Zucker, J.; Baxter, D.; Kumar, N.; Baker, S.; Hurley, J.; Dunlap, J. Prediction of Metabolite Concentrations, Rate Constants and Post-Translational Regulation Using Maximum Entropy-Based Simulations with Application to Central Metabolism of Neurospora crassa. *Processes* **2018**, *6*, 63. [CrossRef]
181. Martyushev, L.M. The maximum entropy production principle: Two basic questions. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 1333–1334. [CrossRef] [PubMed]
182. Vallino, J.J. Ecosystem biogeochemistry considered as a distributed metabolic network ordered by maximum entropy production. *Philos. Trans. R. Soc. B Biol. Sci.* **2010**, *365*, 1417–1427. [CrossRef]
183. Himmelblau, D.M.; Jones, C.R.; Bischoff, K.B. Determination of rate constants for complex kinetics models. *Ind. Eng. Chem. Fundam.* **1967**, *6*, 539–543. [CrossRef]
184. Sorzano, C.O.S.; Vargas, J.; Montano, A.P. A Survey of Dimensionality Reduction Techniques. *arXiv* **2014**, arXiv:1403.2877.
185. Pearson, K. Principal components analysis. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *6*, 559. [CrossRef]
186. Comon, P. Independent component analysis, A new concept? *Signal Process.* **1994**, *36*, 287–314. [CrossRef]
187. Hyv´arinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
188. Andrews, J.G.; Dimakis, A.; Dolecek, L.; Effros, M.; Medard, M.; Milenkovic, O.; Montanari, A.; Vishwanath, S.; Yeh, E.; Berry, R.; et al. A perspective on future research directions in information theory. *arXiv* **2015**, arXiv:1507.05941.
189. Holzinger, A.; H´ortenhuber, M.; Mayer, C.; Bachler, M.; Wassertheurer, S.; Pinho, A.J.; Koslicki, D. On entropy-based data mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*; Springer: Berlin, Germany, 2014; pp. 209–226.
190. Uda, S. Application of information theory in systems biology. *Biophys. Rev.* **2020**, 1–8. [CrossRef] [PubMed]
191. Gohari, A.; Mirmohseni, M.; Nasiri-Kenari, M. Information theory of molecular communication: Directions and challenges. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **2016**, *2*, 120–142. [CrossRef]
192. Navarro, F.C.; Mohsen, H.; Yan, C.; Li, S.; Gu, M.; Meyerson, W.; Gerstein, M. Genomics and data science: An application within an umbrella. *Genome Biol.* **2019**, *20*, 109. [CrossRef] [PubMed]
193. Demchenko, Y.; De Laat, C.; Membrey, P. Defining architecture components of the Big Data Ecosystem. In Proceedings of the 2014 International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, USA, 19–23 May 2014, pp. 104–112.
194. Greene, C.S.; Tan, J.; Ung, M.; Moore, J.H.; Cheng, C. Big data bioinformatics. *J. Cell. Physiol.* **2014**, *229*, 1896–1900. [CrossRef] [PubMed]
195. Hosseini, M.; Pratas, D.; Pinho, A.J. A survey on data compression methods for biological sequences. *Information* **2016**, *7*, 56. [CrossRef]
196. Daily, K.; Rigor, P.; Christley, S.; Xie, X.; Baldi, P. Data structures and compression algorithms for high-throughput sequencing technologies. *BMC Bioinform.* **2010**, *11*, 514. [CrossRef]
197. Xie, X.; Zhou, S.; Guan, J. CoGI: Towards compressing genomes as an image. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 1275–1285. [CrossRef]
198. Ochoa, I.; Hernaez, M.; Weissman, T. iDoComp: a compression scheme for assembled genomes. *Bioinformatics* **2015**, *31*, 626–633. [CrossRef]
199. Mohamed, S.A.; Fahmy, M.M. Binary image compression using efficient partitioning into rectangular regions. *IEEE Trans. Commun.* **1995**, *43*, 1888–1893. [CrossRef]
200. Yu, Y.W.; Daniels, N.M.; Danko, D.C.; Berger, B. Entropy-scaling search of massive biological data. *Cell Syst.* **2015**, *1*, 130–140. [CrossRef]

201. Ishaq, N.; Student, G.; Daniels, N.M. Clustered Hierarchical Entropy-Scaling Search of Astronomical and Biological Data. *arXiv* **2019**, arXiv:1908.08551.
202. Cannon, W.R. Simulating metabolism with statistical thermodynamics. *PLoS ONE* **2014**, *9*, e103582. [CrossRef] [PubMed]
203. Cannon, W.R.; Baker, S.E. Non-steady state mass action dynamics without rate constants: Dynamics of coupled reactions using chemical potentials. *Phys. Biol.* **2017**, *14*, 55003. [CrossRef] [PubMed]
204. Thomas, D.G.; Jaramillo-Riveri, S.; Baxter, D.J.; Cannon, W.R. Comparison of optimal thermodynamic models of the tricarboxylic acid cycle from heterotrophs, cyanobacteria, and green sulfur bacteria. *J. Phys. Chem. B* **2014**, *118*, 14745–14760. [CrossRef]
205. Webb, S. Deep learning for biology. *Nature* **2018**, *554*, 7693. [CrossRef]
206. Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **2016**, *12*, 878. [CrossRef]
207. Wang, Y.; Ribeiro, J.M.L.; Tiwary, P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat. Commun.* **2019**, *10*, 1–8. [CrossRef] [PubMed]
208. Djordjevic, I.B. Quantum Information Theory and Quantum Mechanics-Based Biological Modeling and Biological Channel Capacity Calculation. In *Quantum Biological Information Theory*; Springer: Berlin, Germany, 2016; pp. 143–195.
209. Djordjevic, I.B. Quantum-Mechanical Modeling of Mutations, Aging, Evolution, Tumor, and Cancer Development. In *Quantum Biological Information Theory*; Springer: Berlin, Germany, 2016; pp. 197–236.
210. Djordjevic, I.B. Classical and quantum error-correction coding in genetics. In *Quantum Biological Information Theory*; Springer: Berlin, Germany, 2016; pp. 237–269.