Research Repository UCD

Title	Genetic channel capacity revisited
Authors(s)	Balado, Félix
Publication date	2011-12
Publication information	Balado, Félix. "Genetic Channel Capacity Revisited." Springer, December 2011. https://doi.org/10.1007/978-3-642-32711-7_7.
Conference details	Bionetics 2011, 6th International ICST Conference on Bio-Inspired Models of Network, Information, and Computing Systems, 5-6 December 2011, York, England
Publisher	Springer
Item record/more information	http://hdl.handle.net/10197/3406
Publisher's statement	The final publication is available at www.springerlink.com
Publisher's version (DOI)	10.1007/978-3-642-32711-7_7

Downloaded 2026-05-01 23:36:09

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

Genetic Channel Capacity Revisited

Félix Balado

School of Computer Science and Informatics
University College Dublin
Belfield Campus, Dublin 4, Ireland
Phone: +353 1 7162927
Fax: +353 1 2697262
felix@ucd.ie

Abstract. We revisit previous analyses on the computation of the maximum mutual information between a genetic sequence and its mutated versions down the generations, taking into account the protein translation mechanism of the genetic machinery. This amounts to the application of Shannon's capacity to the study of the transmission of genetic information. Studies on this subject were started by Yockey and then followed by a number of researchers. Here we refine prior analyses employing the Kimura model of base substitution mutations, which is more realistic than the Jukes-Cantor model used by all previous research on this topic. Furthermore we undertake exact computations where prior works just used approximations, and we propose two practical applications of genetic capacity.

1 Introduction

The origin of the application of Shannon's information theoretical concepts to molecular genetics harks back to the work of Quastler [1]. However it was Yockey who led research into this matter for many decades (a compendium of his work can be found in [2]). Yockey has compellingly argued that since information is central to the workings of molecular biology, many aspects of this discipline cannot be fully understood without the help of information theory. More recently Battail [3] has also made a good point of the relevance of information theory for unveiling the workings of evolution.

Information theory relies on probabilistic descriptions of information. For this reason, at the heart of information theoretical explanations of life lie mutations: inescapable random changes undergone by the genomes of organisms, which either go extinct or are accumulated down subsequent generations. As pointed out by Yockey, and by other researchers such as Guiaşu [4], Battail [5], May [6] and Gong et al [7], Shannon's channel capacity [8] is the fundamental limit on how well genetic information can be transmitted in front of mutations. Indeed Shannon's limit applies to any coding strategy —even if produced by evolution by natural selection. Unfortunately the bioinformatics field has paid relatively little attention to date to this potentially important research topic.

Among the aforementioned researchers, only Yockey, Guiaşu, and Gong et al analysed capacity when the protein translation mechanism of the genetic machinery is taken into account, that is, the capacity of the coding regions of genomes. A number of approximations were made in previous works that leave room for further development. In this work we will put prior research in a common framework, and we will refine it by using a more realistic mutation model and by undertaking exact analyses where previous works used approximations. We will also propose two applications of channel capacity in bioinformatics research.

1.1 Notation and Basic Concepts

Calligraphic letters ($\mathcal{X}$) denote sets; $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$. Boldface Roman letters ($\mathbf{x}$) denote row vectors, $\mathbf{x} = [x_1, \cdots, x_N]$. $\mathbf{1}$ is an all-ones vector. Greek capital letters ($\varPi$, $\boldsymbol{\varPi}$) denote matrices, and $(\varPi)_{i,j}$ is the entry of $\varPi$ indexed by (i, j). $(\cdot)^T$ denotes vector or matrix transposition. A Roman letter that appears in uppercase (X) and in lowercase (x) denotes a random variable and a realisation of it, respectively. $p(X = x)$, or just $p(x)$ when unambiguous from the context, is the probability mass function (pmf) or distribution of X. For simplicity, X or the vector $\mathbf{p}_X = [p(X = x)]$ can also denote its distribution depending on the context. $p(X = x|Y = y) = p(x|y)$ denotes a conditional probability. $H(X) = -\sum_{x\in\mathcal{X}} p(x) \log p(x)$ is the entropy of a random variable X with support in $\mathcal{X}$, and $H(X|Y) = -\sum_{y\in\mathcal{Y}} p(y) \sum_{x\in\mathcal{X}} p(x|y) \log p(x|y)$ is the entropy of X conditioned to Y. $I(X; Y) = H(X) - H(X|Y)$ is the mutual information between X and Y. All logarithms are base 2 throughout the paper. The Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ is denoted by $d_H(\mathbf{x}, \mathbf{y})$.

We will summarise next some basic facts about the genetic machinery that we will need in our analysis. The DNA alphabet $\mathcal{X} \triangleq \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$ is formed by the symbols corresponding to its four bases: adenine, cytosine, thymine, and guanine. The nucleotide bases in $\mathcal{X}$ belong to two different chemical categories, namely, purines $\mathcal{R} \triangleq \{\mathrm{A}, \mathrm{G}\}$ or pyrimidines $\mathcal{Y} \triangleq \{\mathrm{C}, \mathrm{T}\}$. A codon —the minimum biologically meaningful "codeword"— is formed by a triplet of consecutive bases in a genetic sequence. A gene is formed by a sequence of codons[1] that can be translated into a sequence of amino acids, which are assembled in the same order imposed by the codons to form a protein. Using their standard short names, the set of amino acids can be written as

$$\mathcal{X}' \triangleq \{\text{Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val, } \mathit{Stp}\}, \tag{1}$$

and therefore $|\mathcal{X}'| = 21$. Every single codon $\mathbf{y} = [y_1, y_2, y_3] \in \mathcal{X}^3$ can be mapped to a unique amino acid or start/stop translation symbol, that is,

$$\xi(\mathbf{y}) = y' \in \mathcal{X}', \tag{2}$$

[1] Our analysis will be independent of the fact that a gene may not always be an unbroken sequence of codons, but rather the intertwining of noncoding sections (introns) with coding sections (exons) in the genomes of eukaryotic cells.

where the mapping $\xi(\cdot) : \mathcal{X}^3 \to \mathcal{X}'$ is established by the nearly-universal *genetic code* (easily found elsewhere, see for instance [9]), which partitions $\mathcal{X}^3$ into $|\mathcal{X}'|$ disjoint subsets of codons. The subset of synonymous codons associated to amino acid $y' \in \mathcal{X}'$ is $\mathcal{S}_{y'} \triangleq \{\mathbf{y} \in \mathcal{X}^3 | \xi(\mathbf{y}) = y'\}$. For instance, with our notation $\mathcal{S}_{\text{Ala}} = \{[\text{G,C,A}], [\text{G,C,C}], [\text{G,C,T}], [\text{G,C,G}]\}$ and $\xi([\text{G,C,A}]) = \text{Ala}$. Note that the ensemble of stop codons is collected under the label *Stp* in (1). *Stp* is loosely classed as an "amino acid" for notational convenience, although it just indicates the end of gene translation and thus does not actually stand for any amino acid. Also Met (and two codons associated to Leu in eukaryotic cells) double as gene translation start symbols. We call the number of codons that map to amino acid y' the *multiplicity* of y'; this is just the cardinality of $\mathcal{S}_{y'}$, that is, $|\mathcal{S}_{y'}|$. Multiplicities are uneven over the set of amino acids, as $|\mathcal{S}_{y'}| \in \{1, 2, 3, 4, 6\}$. Due to the uniqueness of the codon-to-amino acid mapping, $\mathcal{S}_{y'} \cap \mathcal{S}_{w'} = \emptyset$ for $y' \neq w' \in \mathcal{X}'$, and $\sum_{y' \in \mathcal{X}'} |\mathcal{S}_{y'}| = |\mathcal{X}|^3 = 64$ since $\cup_{y' \in \mathcal{X}'} \mathcal{S}_{y'} = \mathcal{X}^3$.

To sum up, the main notational conventions in what follows are that regular Roman letters (i.e. y, Y), bold Roman letters (i.e. $\mathbf{y}$, $\mathbf{Y}$), and primed Roman letters (i.e. y', Y') are associated to bases, codons, and amino acids, respectively. A Greek capital letter which is both bold and primed, for instance $\boldsymbol{\Pi}'$, is a matrix associated to both codons and amino acids.

1.2 Genetic Capacity

As Shannon [8] demonstrated, the maximum mutual information between the input and the output of a channel sets the upper limit on the rate of errorless information transmission for any coding strategy over that channel —crucially, whether the coding strategy is man-made or not. Shannon called this amount *channel capacity*. In molecular genetics contexts we may assimilate mutations — and even natural selection, as we will discuss in Section 2.3— to a probabilistic channel, and we may thus talk of *genetic (channel) capacity*.

The evolution of protein synthesis marks a prominent watershed in the history of life. The mechanism to translate genes into proteins has to be taken into account from that point onwards for the definition of genetic channel capacity to be biologically meaningful. For this reason the relevant changes to be considered in the study of genetic capacity are the changes between a codon $\mathbf{Y} \in \mathcal{X}^3$ in a given DNA sequence and the amino acid $Z'_{(m)} \in \mathcal{X}'$ it translates to after m generations of the corresponding organism, that is, $Z'_{(m)} = \xi(\mathbf{Z}_{(m)})$, where $\xi(\cdot)$ is given by (2) and $\mathbf{Z}_{(m)}$ is the mutated version of $\mathbf{Y}$ after m generations of the organism (see Section 2.2). Alternatively we may also study the genetic capacity between $Y' = \xi(\mathbf{Y})$ and $Z'_{(m)}$. For this reason it is possible to contemplate two different but closely related genetic channels:

1. The *genome-proteome channel*, which is a mutation channel featuring codons at its input and amino acids at its output, and whose capacity is formulated as
$$C' \triangleq \max_{\mathbf{Y}} I(Z'_{(m)}; \mathbf{Y}) \text{ bits/amino acid.} \tag{3}$$

An alternative name for this channel given by Guiaşu is *DNA-to-protein communication channel* [4]. *DNA-mRNA-proteome communication system* was the name used by Yockey [10], who was the first to study (3).

2. The *proteome-proteome channel*, which features amino acids both at its input and at its output and whose capacity is therefore expressed as

$$C'' \triangleq \max_{Y'} I(Z'_{(m)}; Y') \text{ bits/amino acid.} \tag{4}$$

This channel was called *protein communication channel* by Gong et al [7], who were the first to study (4).

1.3 Applications of Genetic Capacity

Previous works on the computation of (3) and (4) did not discuss practical applications of Shannon's capacity in bioinformatics research. Although this is not the main purpose of this work either, we will suggest next two such applications. In the discussion that follows we will use the fact that C'' (and C') must be monotonically nonincreasing on m —as we will prove in Section 3.

1. The information content of a gene is given by its entropy $H(Y')$ (bits/amino acid). This amount can be estimated for a given gene using its empirical distribution of amino acids. According to Shannon's definition of capacity, if $\overline{m}$ is the maximum integer such that $C''|_{m=\overline{m}} = H(Y')$ then the information content of that gene will only be able to survive *unchanged* for at most $\overline{m}$ generations. A looser bound can be obtained with C'. The main consequence is that the genetic capacity establishes a quantifiable "window" of stability for genes. Qualitatively, we can infer that there should exist an evolutionary pressure towards the decrease of $H(Y')$ in successful genes, since this should increase their stability. This may be one reason why genes with uniform Y' do not exist in nature, since uniformity maximises entropy. If the mutation rate were constant and new genes were not created through duplication, a relatively lower value of $H(Y')$ would indicate a relatively older gene.
2. A number of authors have proposed to tackle phylogeny reconstruction by means of empirical estimates of the mutual information between pairs of genes (see for instance [11]). One may establish upper bounds on the branch lengths of the phylogenetic trees thus obtained by means of the genetic capacity, since it must hold that at most $\overline{m}$ generations have elapsed between two genes characterised by Y' and Z' if $\overline{m}$ is the maximum integer such that $C''|_{m=\overline{m}} = I(Z'; Y')$. Finally we must mention that the chain rule for information [12] appears not to have been yet exploited in order to build phylogenetic trees using more than two genes at a time.

2 Mutation Model

Probabilistic mutation models are required in order to evaluate (3) and (4), that is, we need to establish the equivalent of a channel model in digital communications for both cases.

2.1 Low-Level Mutation Model

Our probabilistic models will be built upon a low-level mutation channel describing a channel with bases both at its input and at its output.

Genome-genome channel. We denote the low-level model with this name in order to distinguish it from the high-level models that we will establish in Section 2.2. A simple but realistic model of base substitution mutations (or point mutations) is the Kimura model of molecular evolution [13]. In this model the probability that a base $y \in \mathcal{X}$ mutates to another base $z \in \mathcal{X}$ in the next generation depends on whether the mutation is a transition (that is, an intraclass mutation in which either $z, y \in \mathcal{Y}$ or $z, y \in \mathcal{R}$) or a transversion (that is, an interclass mutation in which $y \in \mathcal{Y}$ and $z \in \mathcal{R}$, or vice versa). The corresponding $|\mathcal{X}| \times |\mathcal{X}|$ base-base transition probability matrix $\Pi \triangleq [p(Z = z|Y = y)]$, with $z, y \in \mathcal{X}$, presents the following structure:

$$\Pi = \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{T} & \text{G} \end{array} \\ \begin{bmatrix} 1-q & \frac{\gamma}{3}q & \frac{\gamma}{3}q & (1-\frac{2\gamma}{3})q \\ \frac{\gamma}{3}q & 1-q & (1-\frac{2\gamma}{3})q & \frac{\gamma}{3}q \\ \frac{\gamma}{3}q & (1-\frac{2\gamma}{3})q & 1-q & \frac{\gamma}{3}q \\ (1-\frac{2\gamma}{3})q & \frac{\gamma}{3}q & \frac{\gamma}{3}q & 1-q \end{bmatrix} \end{array} \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{T} \\ \text{G} \end{array} \tag{5}$$

From this definition, the probability (or rate) of base substitution mutation per generation is just

$$p(Z \neq y|Y = y) = \sum_{z \neq y} p(Z = z|Y = y) = q, \tag{6}$$

for any $y \in \mathcal{X}$. It must hold that $0 \leq \gamma \leq 3/2$ so that all entries of Π are probabilities; in practice $0 < \gamma < 3/2$, since $\gamma = 0$ and $\gamma = 3/2$ would forbid transversion and transition mutations, respectively. The mutation model (5) can incorporate any given transition/transversion ratio ε by setting $\gamma = 3/(2(\varepsilon+1))$. Estimates of ε ranging between 0.89 and 18.67 for RNA and DNA of different organisms are given in [14], which correspond to γ between 0.07 and 0.79. This reflects the fact that transitions are much more likely than transversions, that is, $\varepsilon > 1/2$ virtually always in every organism, and therefore $\gamma < 1$.

The Jukes-Cantor model of molecular evolution, in which all off-diagonal elements of the transition matrix are $q/3$, is obtained as the particular case $\Pi|_{\gamma=1}$ of the Kimura model (5). Notice that this is a *mutation-symmetric model*, since its treatment of mutations between any two different bases is always the same, regardless of the category they belong to (that is, either $\mathcal{R}$ or $\mathcal{Y}$). This less realistic model is the one used in previous computations of (3) and (4) [2, 7]. We will see in Section 3 that the use of the full Kimura model can lead to substantial changes in genetic capacity.

As in prior analyses we will assume that substitution mutations are mutually independent, so that the genetic channel can be taken to be memoryless. When considering m generations of an organism (that is, m cascaded mutation stages) we have a Markov chain $Y \to Z_{(1)} \to Z_{(2)} \to \cdots \to Z_{(m)}$, and model (5) leads

to the overall transition probability matrix $\Pi^m = [p(Z_{(m)} = z|Y = y)]$. This matrix exponentiation can be easily calculated, either directly or by means of diagonalisation techniques.

Note that the model that we are using assumes for simplicity that the base mutation rate (6) stays constant both over the generations and over genetic loci. Although this is the standard practice in this type of analysis it stays a contentious matter. Also, in coding sequences mutation rates vary across the three positions in a codon [9], which will be further discussed in Section 2.3. Our constancy assumption may be cast as a worst-case scenario through a judicious choice of the parameter q. Finally, although substitutions are just one type of mutations, the results that we will obtain using the proposed model constitute an absolute upper bound to genetic capacity when other types of mutations — such as insertions and deletions— are factored in. We must also note that exact capacity analyses of channels with insertions and deletions are still unavailable, even in digital communications settings.

2.2 High-Level Mutation Models

Our high-level models will depend on the extension of the genome-genome channel to describe a channel with codons both at its input and at its output.

Extended genome-genome channel. This channel is modelled by the $|\mathcal{X}|^3 \times |\mathcal{X}|^3$ codon-codon transition probability matrix $\boldsymbol{\Pi} \triangleq [p(\mathbf{Z} = \mathbf{z}|\mathbf{Y} = \mathbf{y})]$ associated to (5), which is just

$$\boldsymbol{\Pi} = \Pi \otimes \Pi \otimes \Pi, \tag{7}$$

where $\otimes$ is the Kronecker product[2]. This is because $p(\mathbf{Z} = \mathbf{z}|\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{3} p(Z_i = z_i|Y_i = y_i)$ according to our independent mutations assumption. Model (7) can be used to characterise any step in the Markov chain $\mathbf{Y} \to \mathbf{Z}_{(1)} \to \mathbf{Z}_{(2)} \to \cdots \to \mathbf{Z}_{(m)}$. Therefore when m such generations are considered the overall transition matrix is $\boldsymbol{\Pi}^m = [p(\mathbf{Z}_{(m)} = \mathbf{z}|\mathbf{Y} = \mathbf{y})]$, which is just $\boldsymbol{\Pi}^m = (\Pi \otimes \Pi \otimes \Pi)^m = \Pi^m \otimes \Pi^m \otimes \Pi^m$ [15].

Genome-proteome channel. This channel is modelled by a $|\mathcal{X}|^3 \times |\mathcal{X}'|$ codon-amino acid transition probability matrix $\boldsymbol{\Pi}' \triangleq [p(Z' = z'|\mathbf{Y} = \mathbf{y})]$. If we define a $|\mathcal{X}|^3 \times |\mathcal{X}'|$ *projection matrix* $\Lambda \triangleq [p(Z' = z'|\mathbf{Z} = \mathbf{z})]$ with entries[3]

$$(\Lambda)_{\mathbf{z},z'} = \begin{cases} 1 & \text{if } \mathbf{z} \in \mathcal{S}_{z'} \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

[2] $\Pi \otimes \Pi = \begin{bmatrix} (\Pi)_{1,1}\Pi & \cdots & (\Pi)_{1,4}\Pi \\ \vdots & \ddots & \vdots \\ (\Pi)_{4,1}\Pi & \cdots & (\Pi)_{4,4}\Pi \end{bmatrix}$

[3] For notational simplicity, a matrix entry is indexed here using a codon-amino acid pair, rather than a pair of integer indices. The amino acid ordering depends on the ordering of $\mathcal{X}'$, whereas the codon ordering depends on the ordering of $\mathcal{X}$. Any such ordering is arbitrary and leads to completely equivalent channels, that is, featuring the same capacity.

then we can simply write the required matrix as

$$\boldsymbol{\Pi}' = \boldsymbol{\Pi}\ \Lambda, \tag{9}$$

where $\boldsymbol{\Pi}$ is given by (7). Of course $\Lambda \mathbf{1}^T = \mathbf{1}^T$, so that $\boldsymbol{\Pi}'$ is properly defined. This transition probability matrix becomes $\boldsymbol{\Pi}'_{(m)} = \boldsymbol{\Pi}^m\ \Lambda$ after m generations.

Proteome-proteome channel. This channel is modelled by a $|\mathcal{X}'|\times|\mathcal{X}'|$ amino acid-amino acid transition probability matrix $\Pi'' \triangleq [p(Z' = z'|Y' = y')]$. If we define a $|\mathcal{X}'| \times |\mathcal{X}'|$ *multiplicity matrix* $\Omega \triangleq \Lambda^T\Lambda$, that is, a diagonal matrix for which $(\Omega)_{z',z'} = |\mathcal{S}_{z'}|$ for all $z' \in \mathcal{X}'$, then Π'' can be written in terms of the codon-amino acid matrix (9) or in terms of the codon-codon matrix (7) as

$$\begin{aligned} \Pi'' &= \Omega^{-1}\Lambda^T\boldsymbol{\Pi}' \\ &= \Omega^{-1}\Lambda^T\boldsymbol{\Pi}\Lambda. \end{aligned} \tag{10}$$

Since $\Lambda^T\mathbf{1}^T = \Omega\mathbf{1}^T$ we can see that Π'' is properly defined. When m generations are considered the transition probability matrix becomes[4] $\Pi''_{(m)} = \Omega^{-1}\Lambda^T\boldsymbol{\Pi}^m\ \Lambda$. See that $\Omega^{-1}\Lambda^T$ is just the $|\mathcal{X}'| \times |\mathcal{X}|^3$ *codon usage matrix* $\Upsilon \triangleq [p(\mathbf{Y} = \mathbf{y}|Y' = y')]$ assuming uniformity. A further increase of C'' could be achieved by optimising Υ; however we will see in Section 3 that this cannot make a great difference since C'' computed using (10) is closely upper bounded by C'.

Finally, while preparing the final version of this article it came to our attention that a similar method for obtaining Π'' from an empirical estimate of Π was previously given in [16], although without the explicit matrix formulation above. In our notation, the approach in [16] uses empirical codon usage to define Υ.

2.3 Natural Selection and Probabilistic Molecular Evolution Models

Coding regions of genomes define proteins and are thus subject to much stronger selection pressures than noncoding regions, whose change can be more obviously considered to be mainly driven by random mutations. Therefore it might appear that a purely probabilistic approach to genetic information transmission would only make sense for noncoding regions, using a low-level model such as (5), whereas it would seem that the effects of natural selection are ignored in the high-level models (9) and (10).

However we must realise that natural selection may also be accounted for in a probabilistic way. This is actually the premise of the so-called *neutral theory of molecular evolution* proposed by Kimura [17]. In order to do so with our high-level models we just need to choose a base mutation rate q that combines the effect of both mutations and natural selection in coding regions. The resulting Π'' matrices, in particular, will be akin to the so-called point accepted mutation matrices (PAM) first estimated by Dayhoff et al [18]. A PAM matrix is basically an empirical estimate of Π'' obtained from real data —unlike our model, without

[4] If Π'' were available but not $\boldsymbol{\Pi}$ one could only use $(\Pi'')^m$ to model this case.

imposing any structure constraint. These matrices are used in the reconstruction of phylogenies and, hence, have to statistically account for evolutive pressures.

We must also remark that, if necessary, the effect of natural selection on coding regions can be reflected even more accurately in our high-level models (9) and (10). The way to do this is to take into account that natural selection implies that the base mutation rate q_w corresponding to the third base in a codon (*wobble* position) is much higher that the base mutation rate q corresponding to any of the bases in the initial duplet. This is because the genetic code implies that mutations affecting the initial duplet of a codon are more prone to induce proteomic change, and hence less likely to survive in subsequent generations. Therefore, if the relevant mutation rates are available, the codon-codon transition probability matrix (7) can be more accurately obtained as $\boldsymbol{\Pi} = \Pi|_q \otimes \Pi|_q \otimes \Pi|_{q_w}$. We will not pursue this idea further in this paper, because we will see in Section 3 that our basic model is sufficient already in order to approximate the results of a PAM model for the purpose of computing genetic capacity.

3 Genetic Capacity Computation

Equipped with the channel models in Section 2.2 we are now ready to obtain the capacities (3) and (4) of the channels that describe the information flow from the genome to the proteome and from the proteome to the proteome, respectively. We will also retrace in this section previous studies on the matter before computing C' and C'' by relying on our Kimura-based models (9) and (10)

In general, the channel matrix $\boldsymbol{\Pi}'_{(m)}$ required to compute (3) does not correspond to a symmetric[5] (or weakly symmetric[6]) channel, which would evince uniform $\mathbf{Y}$ as the maximising input distribution. Similarly the channel defined by $\Pi''_{(m)}$ is not symmetric either, and thus uniform Y' does not necessarily lead to capacity in (4). In spite of this, the required optimisations can be easily undertaken numerically by means of the Blahut-Arimoto algorithm [19]. If $\mathbf{p_Y} = [p(\mathbf{Y} = \mathbf{y})]$ is the pmf that yields C' in (3) observe that we may not always be able to find a pmf $\mathbf{p}_Y = [p(Y = y)]$ such that $\mathbf{p}_Y \otimes \mathbf{p}_Y \otimes \mathbf{p}_Y = \mathbf{p_Y}$. This means that (3) and (4) are upper bounds with respect to maximising the corresponding mutual informations on Y (that is, the input distribution of bases), which is not so straightforward but not our goal here either.

Before any actual computation, one can already see that both C' and C'' have to be monotonically nonincreasing on m. This is because we can establish the Markov chain $Y' \to \mathbf{Y} \to Z'_{(1)} \to \cdots \to Z'_{(m)}$ which implies both that $I(Z'_{(1)}; \mathbf{Y}) \geq \cdots \geq I(Z'_{(m)}; \mathbf{Y})$ and that $I(Z'_{(1)}; Y') \geq \cdots \geq I(Z'_{(m)}; Y')$ by repeatedly applying the data-processing inequality [12]. We can also see

[5] All rows the transition probability matrix are permutations of the same set of probabilities, and so are its columns.

[6] All rows of the transition probability matrix are permutations of the same set of probabilities, and all its columns add up to the same number.

that $C' \geq C''$ for any given m, as the same Markov chain similarly implies that $I(Z'_{(m)};\mathbf{Y}) \geq I(Z'_{(m)};Y')$.

Genome-proteome capacity. As regards prior work on this subject, it is worth mentioning the contribution by Guiaşu [4] in the first place, although this is not the first work dealing with (3). Guiaşu's analysis is for $m = 1$, and uses the transition probability matrix $\boldsymbol{\Pi}'|_{q=0} = \boldsymbol{\Pi}|_{q=0}\ \Lambda = \Lambda$ which corresponds to a mutation-free scenario, that is, $I(Z';\mathbf{Y}) = H(Z') = H(Y')$. The maximum in this case can be analytically obtained by using any distribution of $\mathbf{Y}$ that makes Y' uniform, that is, $\mathbf{p_Y}\Lambda = \mathbf{1}/|\mathcal{X}'|$, for instance $\mathbf{p_Y} = (\mathbf{1}\Omega^{-1}\Lambda^T)/|\mathcal{X}'|$. Guiaşu observed that the corresponding maximum $C'|_{q=0} = \log|\mathcal{X}'| = 4.39$ bits/amino acid is not observed in nature, and he also speculated about (3) when $q > 0$.

This case was actually first tackled by Yockey [10] for $m = 1$ —Guiaşu's work was a later but apparently independent development. Yockey assumed the Jukes-Cantor model of substitution mutations, which he referred to as "white genetic noise", and used an approximation $\widetilde{\boldsymbol{\Pi}}$ of $\boldsymbol{\Pi}|_{\gamma=1}$ in which all entries that correspond to more than one base mutation per codon are nulled. An analytical expression of $\widetilde{\boldsymbol{\Pi}}$ can be obtained as follows. Given the sets of codons $\mathcal{I}_{\mathbf{y}} \triangleq \{\mathbf{z} \in \mathcal{X}^3 | d_H(\mathbf{y},\mathbf{z}) = 1\}$ for $\mathbf{y} \in \mathcal{X}^3$, see that $|\mathcal{I}_{\mathbf{y}}| = (|\mathcal{X}|-1)^3 = 9$ for any $\mathbf{y}$. Therefore Yockey's approximation is

$$(\widetilde{\boldsymbol{\Pi}})_{\mathbf{y},\mathbf{z}} = \begin{cases} \alpha & \text{if } d_H(\mathbf{y},\mathbf{z}) = 1 \\ 1-9\alpha & \text{if } \mathbf{y} = \mathbf{z} \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

with $\alpha = (1-(1-q)^3)/9$. Yockey implicitly approximates this parameter as

$$\alpha \approx \frac{q}{3} \tag{12}$$

which is accurate for $q \ll 1$; however this approximation breaks down for $q > 1/3$ since in this range the diagonal entries of $\widetilde{\boldsymbol{\Pi}}$ must be negative for its rows to add up to one. Yockey discussed how to compute $I(Z';\mathbf{Y})$ using the channel matrix $\widetilde{\boldsymbol{\Pi}}' = \widetilde{\boldsymbol{\Pi}}\ \Lambda$ (which is just Table 5.2 in [2, page 50]) and a given distribution of $\mathbf{Y}$, although he did not attempt maximisation, and he also produced the following approximation [2, page 52]

$$\begin{aligned} I(Z';\mathbf{Y}) \approx\ & H(\mathbf{Y}) - 1.7915 - 9.815\alpha \\ & +34.2108\alpha^2 + 6.8303\alpha\log\alpha. \end{aligned} \tag{13}$$

In Figure 1 we plot C' for $m = 1$ and the whole range of q using Yockey's $\widetilde{\boldsymbol{\Pi}}'$ matrix and the Kimura-based $\boldsymbol{\Pi}'|_{\gamma=1}$ matrix obtained from (9) (that is, Jukes-Cantor based). All results are computed using the Blahut-Arimoto algorithm. The capacity computed with $\widetilde{\boldsymbol{\Pi}}'$ is accurate for $q \ll 1$, but it plateaus as q increases. See that with an exact analysis $C'|_{q=3/4} = 0$ when $\gamma = 1$, since in this case $\boldsymbol{\Pi} = \mathbf{1}^T\mathbf{1}/|\mathcal{X}|^3$ and then no information can be conveyed because any codon can mutate to any other with uniform probability. C' is increasing in the range

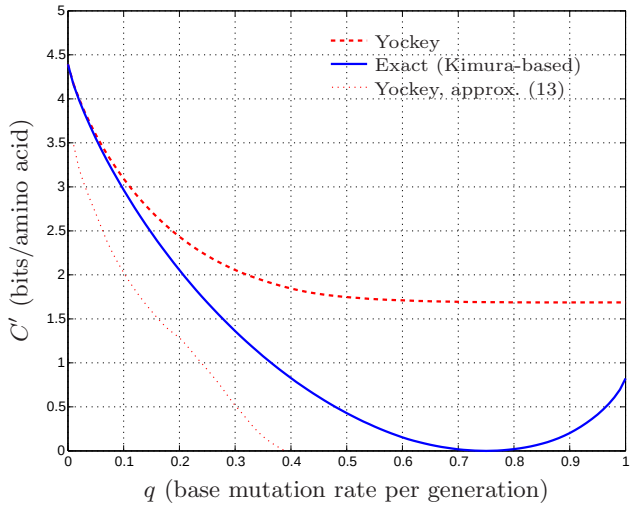

Fig. 1. Genome-proteome genetic capacity ($\gamma = 1$, $m = 1$).

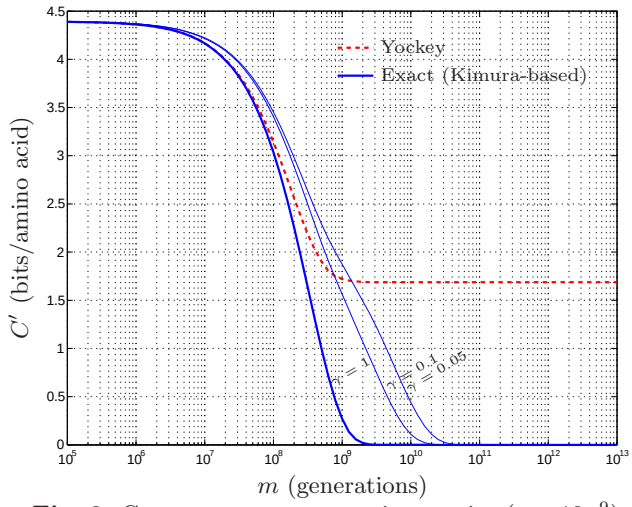

Fig. 2. Genome-proteome genetic capacity ($q = 10^{-9}$).

$q \in [3/4, 1]$ because the uncertainty about $\mathbf{Y}$ given Z' is smaller when $q = 1$ than when $q = 3/4$: given z' when $q = 1$, if a base z is the same for one of the three positions of all $\mathbf{z} \in \mathcal{S}_{z'}$ this implies that $y \neq z$ for the corresponding base in $\mathbf{Y}$ (and hence capacity is higher than when $q = 3/4$). Finally, both plots coincide at $C'|_{q=0}$ with the value given by Guiaşu. As for approximation (13), computed using the optimal distribution of $\mathbf{Y}$ obtained when applying the Blahut-Arimoto algorithm, we can see that it is not very exact: it becomes negative when $q \gtrapprox 0.4$.

Yockey suggested in [2] that his model could be extended to comprise transversions and transitions, but argued that this would not make an important difference. Figure 2, which depicts C' as a function of m, shows that this

assumption was not completely correct. In the realistic case $\gamma < 1$, model $\boldsymbol{\Pi}'_{(m)}$ obtained using (9) yields a noticeable capacity increase afforded by mutation-symmetry breaking. Furthermore Yockey's capacity, computed in Figure 2 using $\widetilde{\boldsymbol{\Pi}}'_{(m)} = \widetilde{\boldsymbol{\Pi}}^m \Lambda$, eventually deviates from the exact capacity for $\gamma = 1$ as m increases; in particular it does not tend to zero as $m \to \infty$, which limits the applicability of approximation (11) even for $q \ll 1$.

Proteome-proteome capacity. As in the previous section we will review and discuss previous research on this matter and compare it with our computations. Gong et al [7], who were the first to study (4), used two different mutation models in their genetic capacity analysis:

1. The first one is parallel to Yockey's. Gong et al propose a Jukes-Cantor model of base substitution mutations and put forward computational reasons to make the exact same approximation as Yockey did, that is, they disregard transitions between codons at Hamming distance greater than one. Consequently their resulting channel matrix ($\mathbf{P}$ in Figure 5 from [7], which we will denote here as $\widetilde{\Pi}''$ for the sake of keeping our notation) is essentially[7] obtained as
$$\widetilde{\Pi}'' = \Omega^{-1} \Lambda^T \widetilde{\boldsymbol{\Pi}} \Lambda, \tag{14}$$
where $\widetilde{\boldsymbol{\Pi}}$ is given by (11). This model is used in [7] to obtain C'' for $m = 1$ as a function of the base mutation rate q.
2. The second one is a point accepted mutation (PAM) matrix [18], that is, an unconstrained estimate of Π'' obtained from real data. This model is used in [7] to obtain C'' as a function of m.

The proteome-proteome channel is not symmetric with any of these two models, and hence Gong et al apply the Blahut-Arimoto algorithm to obtain C''. Unfortunately their $\mathbf{P}$ model suffers from the same shortcomings as Yockey's. Figure 3 shows C'' when $m = 1$ for the whole range of q, both using the approximation $\widetilde{\Pi}''$, with $\alpha = (1-(1-q)^3)/9$, and the exact Kimura-based model $\Pi''|_{\gamma=1}$ obtained from (10). The capacity obtained using $\widetilde{\Pi}''$ is accurate for $q \ll 1$ but plateaus around $C'' = 1.5$ bits/amino acid as q increases, which is not correct for the same reasons as Yockey's result for C' in Figure 1. We must also note that the plot corresponding to Figure 3 in [7] (Figure 6(b), where α is approximated as in (12)) shows C'' only up to $\alpha = 1/9$ (that is, $q = 1/3$) for the reason indicated when discussing the range of validity of (12).

Our next comparison is presented in Figure 4, which gives C'' as a function of m both using $\widetilde{\Pi}''_{(m)} = \Omega^{-1} \Lambda^T \widetilde{\boldsymbol{\Pi}}^m \Lambda$ and $\Pi''_{(m)}$ computed using (10) for different values of γ. The results in this plot are once more parallel to the ones for C' (Figure 2), in particular the mutation-symmetry breaking effect is also observed. We observe again that the genetic capacity does not tend to zero when using approximation $\widetilde{\Pi}''_{(m)}$ as the number of generations m increases.

[7] The channel matrices in [7] are $(|\mathcal{X}'| - 1) \times (|\mathcal{X}'| - 1)$, as they leave *Stp* out.

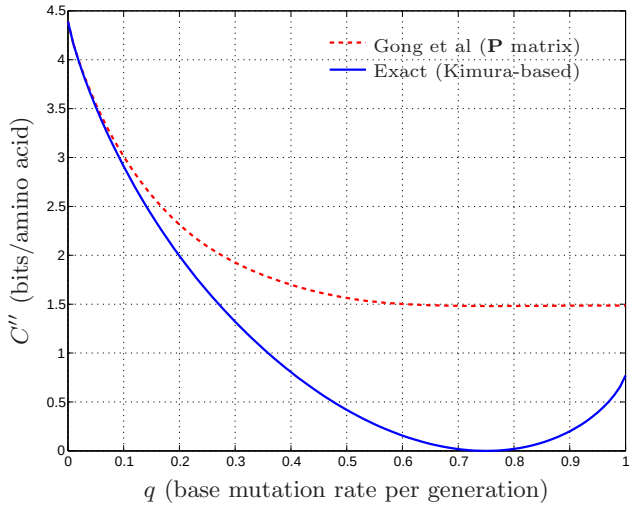

Fig. 3. Proteome-proteome genetic capacity ($\gamma = 1$, $m = 1$).

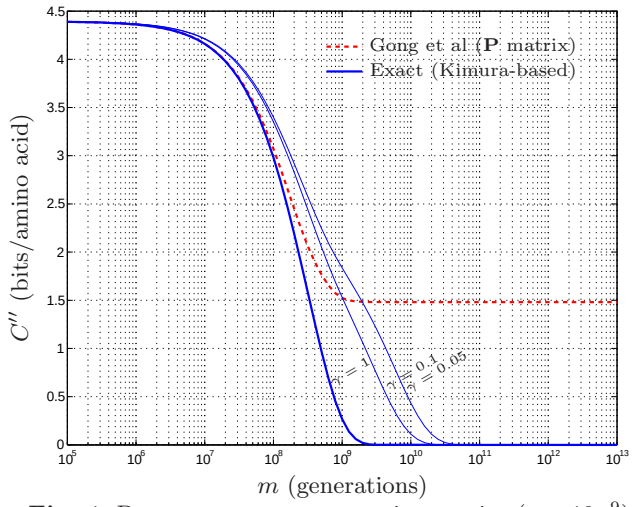

Fig. 4. Proteome-proteome genetic capacity ($q = 10^{-9}$)

This issue of the **P** model is probably behind the use of a PAM model in [7] in order to study C'' as a function of m. With this alternative model these authors able to show that $C'' \to 0$ as $m \to \infty$. However our much simpler Kimura-based model (10) is also able to closely replicate the corresponding result from [7]. To do so we just need to set $q = \frac{1}{3} \times 10^{-2}$ and $\gamma = 1$ so that Π'' represents the 1 PAM matrix given in [18]. Note that $\Pi''_{(m)} \neq (\Pi'')^m$ because $\Lambda\Omega^{-1}\Lambda^T$ is not the identity matrix, but C'' can be empirically shown to be very similar in both cases. The relevance of the comparison presented in Figure 5 is that our model only depends on two parameters, whereas a general PAM matrix requires estimating 380 parameters to model all pairwise amino acid mutabilities.

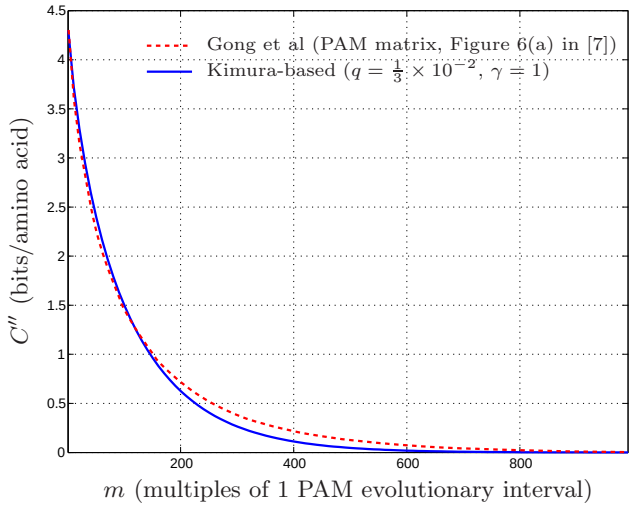

Fig. 5. Proteome-proteome genetic capacity.

4 Conclusions

We have provided a reevaluation of the two genetic capacity measures that take into account the protein translation mechanism, leading to more precise and simpler analyses. We have seen that the channel approximations proposed by Yockey [2] and by Gong et al [7] yield capacity accurately for small values of q (Figures 1 and 3), but even with $q \ll 1$ they are not suitable when accumulated mutations are taken into account as $m \to \infty$ (Figures 2 and 4).

We have also shown that Shannon's capacity turns out to be greater using the more realistic Kimura model with respect to the previously used Jukes-Cantor model. The increase is roughly of one order of magnitude at $C' = 0.01$ ($C'' = 0.01$) in terms of the number of generations m for the realistic transition-transversion parameter $\gamma = 0.1$ (Figures 2 and 4). Finally, we have shown that our simple model of Π'', which depends on two parameters only, is able to closely reproduce the capacity obtained using PAM matrices (Figure 5), which require estimating hundreds of parameters from real data.

Genetic channel capacity remains largely unexplored in the bioinformatics field. We have discussed two concrete applications that could spur further research, but other suitable scenarios can easily be conceived. As a fundamental limit to elementary processes in molecular biology, genetic capacity has the potential to become a staple of practical bioinformatics investigations in the future.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 09/RFP/CMS2212.

References

1. H. Quastler, editor. *Information Theory in Biology*. Urbana: University of Illinois Press, 1953.
2. H. P. Yockey. *Information Theory, Evolution, and the Origin of Life*. Cambridge University Press, 2005.
3. G. Battail. Does information theory explain biological evolution? *Europhys. Lett.*, 40(3):343–348, 1997.
4. S. Guiaşu. *Information Theory with Applications*. McGraw-Hill, 1977.
5. G. Battail. Information theory and error-correcting codes in genetics and biological evolution. In M. Barbieri, editor, *Introduction to Biosemiotics*. Springer, 2007.
6. E. E. May. Bits and bases: An analysis of genetic information paradigms. In *41st Asilomar Conf. on Signals, Systems and Computers (ACSSC)*, pages 165–169, Asilomar, USA, November 2007.
7. L. Gong, N. Bouaynaya, and D. Schonfeld. Information-theoretic model of evolution over protein communication channel. *IEEE/ACM Trans. on Comp. Biol. and Bioinformat.*, 8(1):143–151, January-February 2011.
8. C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423 and 623–656, July and October 1948.
9. W. Li. *Molecular Evolution*. Sinauer Associates, 1997.
10. H. P. Yockey. An application of information theory to the central dogma and the sequence hypothesis. *J. Theor. Biol.*, (46):369–406, 1974.
11. Z. Yu, Z. Mao, L-Q. Zhou, and V. Anh. A mutual information based sequence distance for vertebrate phylogeny using complete mitochondrial genomes. In *Procs. of the IEEE 3rd Intl. Conf. on Natural Computation*, pages 253–257, Haikou, China, 2007.
12. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.
13. M. Kimura. A simple method for estimating evolutionary rate in a finite population due to mutational production of neutral and nearly neutral base substitution through comparative studies of nucleotide sequences. *J. Mol. Biol.*, 16:111–120, 1980.
14. A. Purvis and L. Bromham. Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny. *J. of Mol. Evol.*, 44:112–119, 1997.
15. J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 3rd edition, 1999.
16. P. Mackiewicz, P. Biecek, D. Mackiewicz, J. Kiraga, K. Baczkowski, M. Sobczynski, and S. Cebrat. Optimisation of asymmetric mutational pressure and selection pressure around the universal genetic code. In *Procs. of the 8th Intl. Conf. on Computational Science*, volume 5101 of *Lecture Notes in Computer Science*, pages 100–109, Krakow, Poland, June 2008.
17. M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.
18. M. Dayhoff, R. Schwartz, and B. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(3):345–352, 1978.
19. R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. on Inf. Theory*, 18(4):460 – 473, July 1972.