PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Check for updates

Article submitted to journal

Subject Areas:
Proteomics, Information theory, Software, Music, Text analysis

Keywords:
Conservation, Hartley-Shannon information, protein lengths, software lengths

Author for correspondence:
Les Hatton
e-mail: lesh@oakcomp.co.uk

Are all discrete systems shaped by the same conservation principle ?

Les Hatton[1], Gregory Warr[2]

[1]Faculty of Science, Engineering and Computing, Kingston University, Penrhyn Road, Kingston upon Thames, KT1 2EE, United Kingdom.
[2]Medical University of South Carolina, 96 Jonathan Lucas St, Charleston, SC 29425, USA.

Supplementary Materials.

THE ROYAL SOCIETY
PUBLISHING

Notes

These Supplementary Materials consist of a series of Appendices with more detailed proofs of the results described in the main text. Note that we have left references in to enable peer review but wherever possible, these also appear in the main text to provide the cited authors appropriate recognition.

Appendix A: The Conservation of Hartley-Shannon Information: From Statistical Mechanics to the Canonical CoHSI Distribution

Statistical mechanics is a methodology for predicting component distributions of general systems made from discrete pieces or components subject to restrictions known as constraints. Such systems include gases (made from molecules), proteins (made from amino acids), software (made from programming language tokens) and even boxes containing beads. Conventionally, constraints are applied by fixing the total number of pieces and/or the total energy [1].

Statistical Mechanics: Classical use in physics

To illustrate the method, we describe a classical problem of determining the most likely distribution of particles amongst energy levels.

To see this, the following variational methodology is borrowed from the world of statistical physics, ([2] (p.217-); and for an excellent introduction, see [1]). In the kinetic theory of gases, a standard application is to find the most common arrangement of molecules amongst energy levels in a gas subject to various constraints such as a fixed total number of molecules and fixed total energy. For this, imagine that there are M energy levels, where the number of particles with energy level ε_i is $t_i, i = 1, .., M$.

For this system, the total number of ways W of organising the particles amongst the M energy levels is given by:-

$$W = \frac{T!}{t_1!t_2!..t_M!}, \tag{0.1}$$

where

$$T = \sum_{i=1}^{M} t_i \tag{0.2}$$

The total amount of energy in this system is just the sum of all the particle energies and is given by

$$E = \sum_{i=1}^{M} t_i \varepsilon_i \tag{0.3}$$

In a physical system, E corresponds to the total internal energy and the variational method to follow constrains this value to be fixed; i.e. solutions are sought in which energy is conserved.

Using the method of Lagrangian multipliers and Stirling's approximation as described in [1], will give the most likely distribution satisfying equation (0.1) subject to the constraints in equations (0.2) and (0.3). This is equivalent to maximising the following variational derived by taking the natural log of (0.1). Just as in maximum likelihood theory, taking the log dramatically simplifies the proceedings, in this case the factorials, and allows the use of Stirling's theorem for large numbers. Also, since it is monotonic, a maximum in log W is coincident with a maximum in W. This leads to

$$\log W = T\log T - \sum_{i=1}^{M} t_i \log(t_i) + \lambda\{T - \sum_{i=1}^{M} t_i\} + \beta\{U - \sum_{i=1}^{M} t_i\varepsilon_i\} \tag{0.4}$$

where λ and β are the multipliers [1]. In essence, the variational process envisages varying only the contents t_i of each of the components until a maximum of $\log W$ is found. The maximum is indicated by taking $\delta(\log W) = 0$, (analogous to finding maxima in differential calculus). Noting that

- the variational operator δ acts on pure constants such as $T\log T$, λT and βU to produce zero just as when differentiating a constant,
- the product rule of differentiation gives $\delta(t_i \log(t_i)) = \delta t_i \log(t_i) + t_i \delta(\log(t_i) = \delta t_i(1 + \log(t_i))$,
- ε_i is independent of the variation by assumption,
- T and the t_i are $\gg 1$ (to satisfy Stirling's theorem, although it is surprisingly accurate even for relatively small values).

This leads to

$$0 = -\sum_{i=1}^{M} \delta t_i\{\log(t_i) + \alpha + \beta\varepsilon_i\} \tag{0.5}$$

where $\alpha = 1 + \lambda$. (Further elaboration of this standard technique can be found in Glazer and Wark [1].)

Finally, (0.5) must be true for all variations to the occupancies δt_i and therefore implies

$$\log(t_i) = -\alpha - \beta\varepsilon_i \tag{0.6}$$

for all i.

Using equation (0.2) to replace α, this can be manipulated into the most likely, i.e. the equilibrium distribution, of particles amongst the M components.

$$t_i = \frac{Te^{-\beta\varepsilon_i}}{\sum_{i=1}^{M} e^{-\beta\varepsilon_i}} \tag{0.7}$$

Defining $p_i = \frac{t_i}{T}$ means that p_i can be interpreted as a probability density function since it is non-negative everywhere and its sum everywhere is equal to 1. Then (0.7) yields

$$p_i = \frac{e^{-\beta\varepsilon_i}}{\sum_{i=1}^{M} e^{-\beta\varepsilon_i}} \tag{0.8}$$

This is a result with a profound interpretation in physics. It states

If we consider all possible systems which share the same number of particles T and the same total energy U (i.e. energy is *conserved*), then given any one example of such a system, it is overwhelmingly likely to obey (0.8). In other words by considering all possible systems with these parameters and constraining them to have the same T and U, probability distribution (0.8) is overwhelmingly likely, provided the t_i are large enough for Stirling's approximation to hold.

In this case, the distribution is exponential and exactly as is found in nature - *exponentially fewer particles occupy higher energy levels.* All the above has been known for decades and is extremely successful at explaining classical systems such as gases and even quantum mechanical systems; the methodology of statistical mechanics however is exceedingly versatile, so let us consider a simple model just consisting of boxes of coloured beads.

Conservation of Hartley-Shannon Information

Identical boxes of beads

Let us put some flesh on the meaning of "overwhelmingly likely" as used earlier in this paper. Consider now a system of M boxes of identical beads, where the i^{th} box contains t_i beads and M is reasonably large. If we have T beads in total numbered by sequence, where $T = \sum_{i=1}^{M} t_i$, so that they are distinguishable by their order, the number of possible ways of arranging them in each of the M boxes is given by

$$\Omega = \frac{T!}{\prod_{i=1}^{M}(t_i!)} \tag{0.9}$$

Suppose there are $M = 10$ boxes and $T = 100$ beads and we simply assign them one by one to a randomly chosen box. We would be very surprised if the first box contained all the beads with the others empty, and the number of ways this can happen according to (0.9) is $100!/(100! \times 0! \times 0!..0!) = 1$. If each box contains 10 beads however as shown in Fig. 1, this situation can happen in $100!/(10! \times 10! \times ..10!)$ ways, which is approximately 10^{100}, a gigantic number.

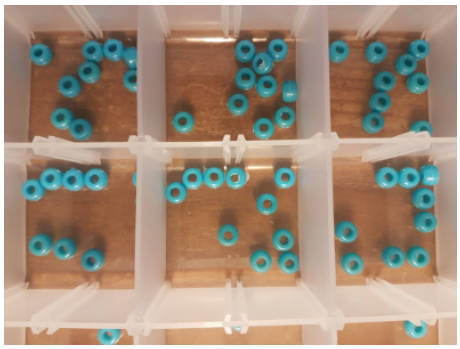

Figure 1: Boxes containing exactly the same number of exactly the same bead.

In other words, we are overwhelmingly more likely to see equal box populations than 1 single filled box. In fact statistical mechanics allows us to prove that, in this case, equal population is by far the most likely distribution of contents simply by finding the maximum of (0.9) subject to a fixed number of T beads in the form

$$\log \Omega = T \log T - T - \sum_{i=1}^{M} \{t_i \log(t_i) - t_i\} + \alpha\{T - \sum_{i=1}^{M} t_i\} \tag{0.10}$$

where the constraint on fixing T is controlled by the Lagrangian parameter α. Finding the maximum of (0.10) using the standard $\delta()$ method [1], gives the solution $t_i \sim constant$, corresponding to equal box populations. Frank also demonstrates this in his maximum entropy formulation [3], p. 9.

Heterogeneous boxes of beads

We now describe an extension which is directly relevant to systems such as the known proteome or computer programs. Consider Fig. 2.

Figure 2: The heterogeneous case where each box contains mixed types and different numbers of beads. This is relevant to proteins, computer program functions and the length distribution of words in texts.

Here the boxes contain differently coloured beads. We envisage this as the i^{th} box containing t_i beads selected randomly from a *unique alphabet* of a_i colours, ordered by sequence. For the proteome, the "colours" correspond to different amino acids and for software functions they correspond to different programming language tokens.

Following the justification given in the main text, we utilise the great flexibility of statistical mechanics by generalizing the payload to be Hartley-Shannon (H-S) information content instead of energy [4].

The H-S information content of the i^{th} box I_i, (Appendix p. 24) is simply the log of the number of ways of arranging the beads in that box, *so that it is guaranteed to contain at least one of each of the a_i colours*. H-S information is completely agnostic about what the colours actually mean, *indeed Hartley specifically advised against attaching any meaning to a token [5].* For maximum generality, this is precisely the behaviour we seek. The only thing that matters is that beads *change* colour, so the actual colour is irrelevant and the total H-S information is just the sum of the information for each box. Presented with such a system, we can ask what is the most likely distribution of contents for systems for which both the total number of beads and the total H-S information are conserved ? We must also recall that proteins and software are both constructed sequentially so we are considering systems where beads are distinguishable by the order in which they appear, but the actual order is irrelevant.

The relevant variational form we must solve is therefore

$$\log \Omega = T \log T - T - \sum_{i=1}^{M} \{t_i \log(t_i) - t_i\} + \alpha\{T - \sum_{i=1}^{M} t_i\} + \beta\{I - \sum_{i=1}^{M} I_i\} \quad (0.11)$$

The only term which is different in this formulation from the classical solution derived above (0.4), is the last term on the right hand side of (0.11). In the variational methodology, each term has the $\delta()$ operation applied in order to vary the t_i and derive the distribution in (0.5), so we are interested specifically in

$$\delta\big(\beta\{I - \sum_{i=1}^{M} I_i\}\big) = -\beta \sum_{i=1}^{M} \delta(I_i) = -\beta \sum_{i=1}^{M} \frac{dI_i}{dt_i} \delta t_i, \quad (0.12)$$

since I is being held constant.

Now consider what happens when boxes are very large compared with their unique alphabet, i.e. $t_i \gg a_i$. In this case, [4], the information content is

$$I_i = \log(a_i \times a_i \times \ldots \times a_i) = \log(a_i^{t_i}) = t_i \log a_i \tag{0.13}$$

In other words, we select t_i times from a choice of a_i colours secure in the knowledge that since $t_i \gg a_i$, it is very unlikely that any of the a_i colours would be missed out and we therefore meet the requirement of having *exactly* a_i unique colours.

In this case, (0.12) becomes

$$-\beta \sum_{i=1}^{M} \frac{dI_i}{dt_i} \delta t_i = -\beta \sum_{i=1}^{M} \frac{d(t_i \log a_i)}{dt_i} \delta t_i = -\beta \sum_{i=1}^{M} (\log a_i) \delta t_i \tag{0.14}$$

(0.14) fits perfectly into the variational methodology leading to (0.11), modifying (0.6) to give

$$\log(t_i) = -\alpha - \beta \log a_i \tag{0.15}$$

The analogue of (0.8) is therefore

$$p_i \equiv \frac{t_i}{T} = \frac{a_i^{-\beta}}{\sum_{i=1}^{M} a_i^{-\beta}} \tag{0.16}$$

To summarize, maximising (0.9) subject to a fixed total number of beads T AND a fixed total H-S information $I = \sum_{i=1}^{M} I_i$ is directly analogous to maximising (0.4) with $\log a_i$ replacing ϵ_i. Like its classical equivalent (0.8), (0.16) is also fundamental. It states

In any discrete system satisfying the model described here, the tail (i.e. $t_i \gg a_i$) of the distribution of unique alphabets is overwhelmingly likely to obey a power-law.

Note that by analogy with (0.4), we can interpret $t_i \log a_i$ as each bead carrying a payload of $\log a_i$, so that even though H-S information is token agnostic, the beads in a particular box still carry a box-dependent payload which is a function of the unique alphabet of colours in that box, a_i. This is exactly analogous to $t_i \varepsilon_i$ being interpreted as each particle carrying an energy ε_i in classical statistical mechanics. In other words, each box behaves as if it had a fixed *information* level $\log a_i$ determined by its unique alphabet. In a protein for example, this has the intriguing implication that even though H-S information is token-agnostic, a particular amino acid in one protein may carry a different information payload than when present in another protein, simply because its neighbours are different.

The asymptotic dual distribution

As pointed out by [6], (0.16) has a dual solution. With some algebra, it can be shown that

$$q_i \equiv \frac{a_i}{A} = \frac{t_i^{-1/\beta}}{\sum_{i=1}^{M} t_i^{-1/\beta}}, \tag{0.17}$$

where

$$A = \sum_{i=1}^{M} a_i \tag{0.18}$$

Note here that A emerges naturally as the sum of the unique alphabets of each component. It is *not* the size of the unique alphabet across all components. This is simply another manifestation of the token-agnosticism of Hartley-Shannon information - system-wide uniqueness of the alphabet simply does not emerge as a requirement. The only requirements for a pdf are that it be positive

definite and normalisable so this in no way detracts from the fact that (0.17) is also a power-law. In other words,

In any discrete system satisfying the model described here, for example the proteome or software functions, the tail (i.e. $t_i \gg a_i$) of the length distribution is overwhelmingly likely to obey a power-law.

Note also the natural appearance of the reciprocal slope $1/\beta$. This value is not found in the datasets here but this difference is discussed and we think resolved in the discussion of alphabets in music later.

The chocolate box analogy and additive partitions: the CoHSI distribution

For smaller boxes containing fewer beads, the above value of I_i (0.13) is not correct. If t_i is closer in size to a_i, (it cannot be smaller since the length must be at least equal to the unique alphabet), there is an increasingly high probability that we might miss out one of the colours in the unique alphabet as we select our t_i beads, breaking the fundamental assumption that each box contains a unique alphabet of **exactly** a_i. We must therefore make different provisions as the boxes get smaller.

Figure 3: A box of 22 chocolates chosen from 12 different types as shown on the left.

The situation is akin to boxes of mixed chocolates, Fig. 3. Such boxes are constructed from a fixed set of chocolates advertised on the lid, and every box must contain at least one of each. Larger boxes simply contain more than one of some kinds. In how many ways can such boxes be created ?

Note that it is simple to find an algorithm to guarantee that the unique alphabet is exactly a_i. All that is necessary is to fill any a_i places with one chocolate of each type and then fill the remaining $t_i - a_i$ at random from the available types. The number of ways of doing this is

$$\left((a_i!).({}^{t_i}C_{a_i})\right).\left(a_i^{(t_i-a_i)}\right), \tag{0.19}$$

where ${}^{n}C_r = n!/((n-r)!r!)$ is the combination operator. This however, is not the same as counting *all* the possible ways of filling the box such that it contains exactly a_i unique chocolates.

We are trying to find the number of different ways of filling the i^{th} box with t_i chocolates chosen from a set of exactly a_i unique chocolates such that the box contains at least one of each, and we must do this if possible in a way which fits into the statistical mechanical framework so we can use its methodology.

To explore this, suppose we have a box of $t_i = 5$ chocolates such that it contains *exactly* $a_i = 2$ different chocolates of types A and B. The total number of ways this can be done $N(t_i, a_i)$, is given by

$$N(5,2) = \frac{5!}{1!4!} + \frac{5!}{4!1!} + \frac{5!}{3!2!} + \frac{5!}{2!3!} \tag{0.20}$$

Note

- The first term on the right hand side of (0.20) is the total number of ways of selecting 5 chocolates by using 1 chocolate of type A and 4 chocolates of type B. This is equal to 5 (ABBBB, BABBB, BBABB, BBBAB, BBBBA).
- The second term corresponds to 4 chocolates of type A and 1 of B and is also equal to 5 (BAAAA, ABAAA, AABAA, AAABA, AAAAB).
- The third term corresponds to taking 3 of type A and 2 of type B. This is equal to 10, (AAABB, AABAB, AABBA, ABAAB, ABABA, ABBAA, BBAAA, BABAA, BAABA, BAAAB).
- The fourth term corresponds to taking 2 of type A and 3 of type B. This is also equal to 10, (BBBAA, BBABA, BBAAB, BABBA, BABAB, BAABB, AABBB, ABABB, ABBAB, ABBBA).

There are no other ways of arranging the box such that there are exactly 2 kinds of chocolate and exactly 5 chocolates altogether. There are therefore 5 + 5 + 10 + 10 = 30 different such boxes in total. Note that (0.19) gives $(2!).({}^5C_2)).(2^{(5-2)}) = 160$ boxes. This over-counting is because a box such as ABBAB could be generated several times by that algorithm, for example, by filling the first two places with AB and then the rest at random or by filling the first and third places with AB and the rest at random.).

The denominators of (0.20) correspond to elements of the *additive compositions*[1] of size 2 of the number 5. These are

$$5 = 1 + 4; 5 = 4 + 1; 5 = 3 + 2; 5 = 2 + 3 \tag{0.21}$$

There are other additive compositions such as $2 + 2 + 1$, but this corresponds to three different kinds of chocolate so must be excluded.

The fact that the compositions are *additive* presents a real complication when merging with the methodology of statistical mechanics because it breaks the steps leading from (0.12)-(0.16) by introducing the log of a definition which includes additive terms. Prior to discovery of the recursive method which follows, the solution was simply trapped between a lower and upper bound. The lower bound consisted of just one of the terms leading to the recursive definition and the upper bound was the pure power-law (0.16). The recursive method is however far more compelling.

First we slightly modify the definition in (0.20) by letting $N(t_i, a_i; a_i')$ be the number of ways of producing a chocolate box with t_i chocolates containing exactly a_i' unique types chosen from a total unique number of types of a_i. In this notation, for example, $N(5, 2; 1) = 2$ and $N(5, 2; 2) = 30$. The distinction between a_i and a_i' is to make way for the use of recursion.

It can be verified that the following recursion then generates the desired total number of ways $N(t_i, a_i; a_i)$ of generating a chocolate box of t_i chocolates from a unique set of chocolates a_i.

for $t_i = 1, .., t_i(MAX)$ **do**
  for $a_i = 1, .., t_i$ **do**
    $N(t_i, 1; 1) = 1;$
    for $i = 1, .., (a_i - 1)$ **do**
      $N(t_i, a_i; i) \leftarrow {}^{a_i}C_i N(t_i, i; i)$
    end for
    $N(t_i, a_i; a_i) \leftarrow a_i^{t_i} - \sum_{i=1}^{a_i - 1} N(t_i, a_i; i)$
  end for

[1] https://en.wikipedia.org/wiki/Partition_(number_theory), accessed 02-Jun-2017.

end for

The corresponding Hartley-Shannon information content for a box containing t_i chocolates chosen from a unique alphabet of a_i chocolates is therefore given by

$$I_i = \log\left(N(t_i, a_i; a_i)\right) \tag{0.22}$$

In contrast, the equivalent form for the pure power-law (0.13) is

$$I_i = \log\left(a_i^{t_i}\right) = t_i \log a_i \tag{0.23}$$

As we have seen, applying the $\delta()$ operator to (0.23) as shown in (0.14) then leads to the pure power-law equation

$$\log t_i = -\alpha - \beta(\log a_i), \tag{0.24}$$

whereas applying the $\delta()$ operator to (0.11) using (0.12) and (0.22), leads to the canonical equation

$$\log t_i = -\alpha - \beta(\frac{d}{dt_i} \log N(t_i, a_i; a_i)), \tag{0.25}$$

We note in passing that this uses the simplest form of Stirling's approximation to $\log(t_i!) \approx t_i \log t_i - t_i$. We also experimented with using Ramanujan's form [7] in which case we get

$$\log t_i + \frac{1 + 8t_i + 24t_i^2}{6(t_i + 4t_i^2 + 8t_i^3)} = -\alpha - \beta(\frac{d}{dt_i} \log N(t_i, a_i; a_i)), \tag{0.26}$$

This however, has only a small effect on the resulting pdf so we will continue to discuss the solution in terms of (0.25).

We can now see the problem posed by (0.25). The presence of the recursive definition of $N(t_i, a_i; a_i)$ prevents the clean separation of factors by the log operation. This we must solve computationally allowing for the difficulties caused by the large factorial values which arise even for modest values of (t_i, a_i).

Here, the unique alphabet a_i is playing a dual role as the frequency in a pdf by analogy with (0.16) using (0.17) and *(0.25) implicitly defines the length distribution at all scales of heterogeneous discrete systems. We will refer to it as the canonical CoHSI distribution for heterogeneous systems.*

Note finally that following the argument that led up to (0.13), for $t_i \gg a_i$, we are guaranteed that

$$\log N(t_i, a_i; a_i) \to t_i \log a_i, \tag{0.27}$$

so the full solution does indeed correctly asymptotes to the pure power-law. This will be confirmed during the computation with both forms being displayed together.

(0.25) defines the canonical implicit pdf with solutions (t_i, a_i) which a) conserves H-S information and b) asymptotes to the pure power-law pdf (0.24) for $t_i \gg a_i$ as required for any heterogeneous system at all scales.

Computational aspects of the CoHSI distribution

Before we proceed with this, there is a technical limitation to overcome since (0.25) is implicit. As we pointed out in the main text, there is a precedent for this in the definition of Tsallis entropy [8,9], although in our case, the implicit nature of the pdf arises naturally from CoHSI. (In Tsallis entropy, the entropy term is adjusted using an additional parameter and this adjustment can lead to an implicit pdf.)

We must therefore generalise the argument from integer values of (t_i, a_i), to the real line. This will not affect our computation of factorials in the recursive evaluation of $N(t_i, a_i; a_i)$ however, which are done at integer values of t_i, a_i, with interpolation for non-integer values.

We solved (0.25) using the following procedure.

(i) Compute a discrete grid of values of $N(t_i, a_i; a_i)$ for those integer values of t_i, a_i where there is no overflow due to the large factorials. In a normal perl program, we were able to do this for the grid $a_i = 1, 2, .., 30, t_i = 1, 2, .., 300$.
(ii) Compute the derivative of this grid $d/dt_i(N(t_i, a_i; a_i))$ using a second-order difference approximation.
(iii) Solve (0.25) interpolating $d/dt_i(N(t_i, a_i; a_i))$ as necessary whilst computing the pure power law solution (0.24) concurrently as a check.

The full CoHSI pdf - the solution of (0.25), and the pure power-law solution - the solution of (0.24), using the same parameters of $\alpha = 4, \beta = 0.8$, are shown together at two different scales as Figs. 4a and 4b. Both curves are normalised such that the area under the full CoHSI pdf integrates to 1 as required. The shaded zone corresponds to the region where the full solution departs from the pure power-law solution. Note that the numerical approximation for the differential and the simple linear interpolation used means that the first few points for the full CoHSI solution are not reliably calculated and are not shown. The behaviour is clear from the remaining points however and both Figs. 4a and 4b are within the range in which the grid of t_i, a_i values could be computed demonstrating the rapid convergence of the CoHSI solution (0.25) to the pure power-law solution (0.24) from about $t_i = 20$ onwards. For $t_i > 300$, the power-law solution (0.24) is used. Full details and software are included in the deliverability package accompanying this paper.

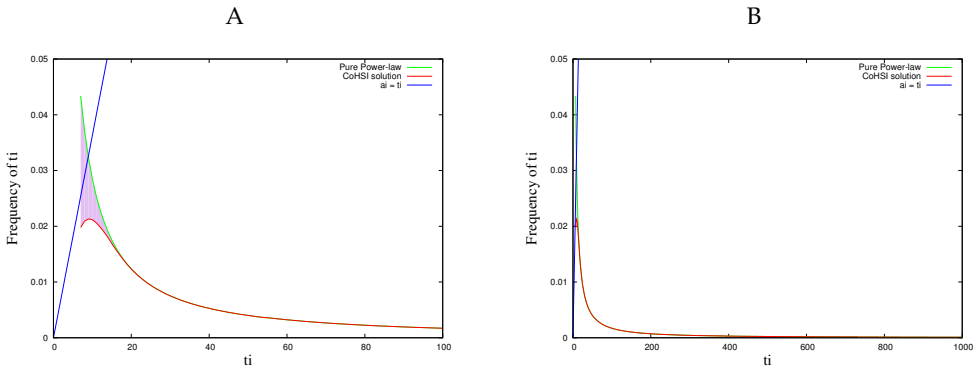

Figure 4: The length distributions using the same modelling parameters $\alpha = 4, \beta = 0.8$ for (A) the full CoHSI solution (0.25), the pure (asymptotic) power-law (0.24) and the boundary condition $t_i = a_i$ for components smaller than 100 tokens and (B), the same data for components up to 2,000 tokens long. Note that no real solutions can exist to the left of $t_i = a_i$.

We make the following observations about Figs. 4a and 4b with respect to (0.25) and (0.24).

- The qualitative behaviour of the full CoHSI solution around the unimodal peak remains sharp but is more rounded than the pure power-law solution and is qualitatively similar to the software data close up of the main text and repeated here as Fig. 5. The sharpness of the peak is related to values of α, β as we shall see shortly but we also note the boundary condition naturally emerging in this theory that $t_i \geq a_i$, i.e., no component can be shorter than its unique alphabet.

- The full CoHSI solution naturally asymptotes to the pure power-law behaviour as required obviating the need to juxtapose dual regions, one matched by a lognormal distribution and the other by a power-law has been used in the past, [10,11]. In our theory, this transition emerges completely naturally as the implicit solution of (0.25).

We can compare the behaviour around the peak with a close-up of the dataset of Fig. 1b of the main text, as shown as Fig. 5. Even on observed data, the transition from power-law to near linearity is abrupt, taking place over perhaps 10 tokens, and is qualitatively very similar to the full CoHSI solution.

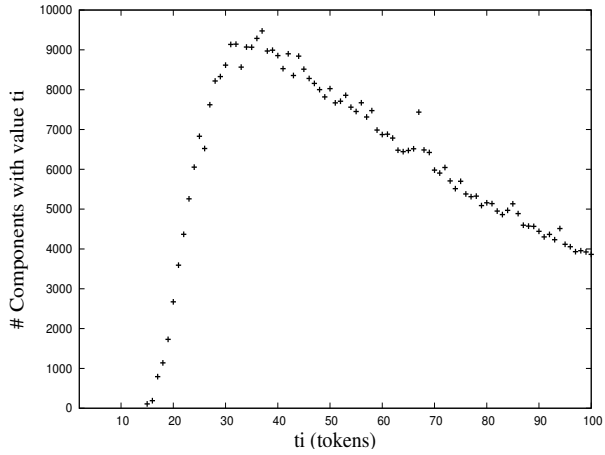

Figure 5: A close-up around the peak of the measured dataset shown as Fig. 1b of the main text.

We close this discussion by investigating solutions of (0.25) for different values of the undetermined Lagrange parameters α, β, normalised such that the areas under each curve are 1 as required for a pdf so that they can be compared.

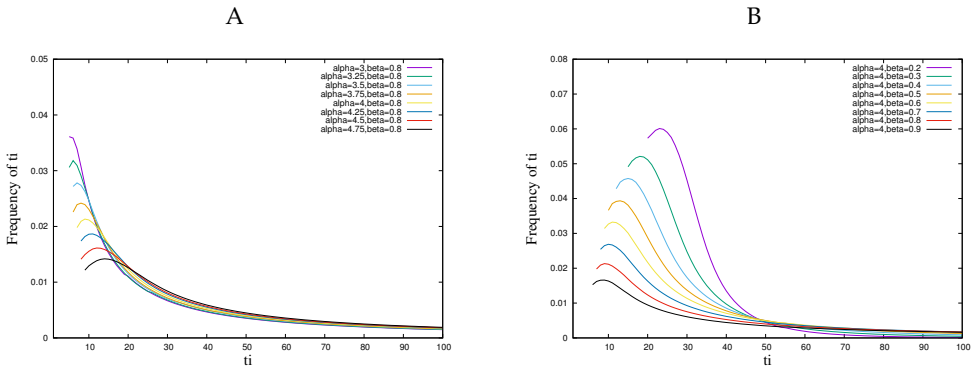

Figure 6: Solutions of the full CoHSI equation (0.25) (A) for various values of α for fixed β and (B), for various values of β for fixed α. The graphs have $\alpha = 4, \beta = 0.8$ in common.

We make the following observations about Figs. 6a and 6b.

- For fixed β, as α (the normalisation parameter) decreases, the position of the peak moves slowly towards lower values of t_i, whilst the amplitude of the peak increases and becomes more sharply defined. The power-law nature for larger t_i does not change much.
- For fixed α, as β (the power-law parameter) decreases, the position of the peak moves slowly towards higher values of t_i, whilst the amplitude of the peak increases. The sharpness of the peak does not change much. As β decreases, the power-law approaches zero for larger t_i more quickly.

To summarise, these results strongly support the thesis of this paper that the Conservation of Hartley-Shannon Information (CoHSI) acts as a constraint on how the length and alphabet size distributions of systems of a given size T and total Hartley-Shannon information I, can evolve at all scales giving an excellent qualitative match with the observed distribution Fig. 5, and which does not require juxtaposing existing pdfs of known properties.

Approximate properties of the heterogeneous CoHSI distribution

From the theory which led up to Figs. 4a, 4b, we can approximate the distribution satisfactorily by glueing together a right-angled triangle up to the modal value a_{max} at $t = t_{max}$, say and a power-law afterward because the solution corresponding to (0.22) transitions from power-law to almost linear behaviour so quickly. In other words, we can define the approximate canonical distribution $c(t)$ as follows

$$c(t) = \begin{cases} (\frac{2(\gamma-1)}{(t_{max}^2(\gamma+1))})t & 0 \le t \le t_{max} \\ (\frac{2(\gamma-1)}{(t_{max}(\gamma+1))})(\frac{t}{t_{max}})^{-\gamma} & t_{max} < t < \infty \end{cases}$$

We require $\gamma > 1$ for this to be positive definite. This has been normalised so as to integrate to unity over it's support $[0, \infty]$. This approximation will allow us to make useful inferences. First we will calculate the mean location and spread of this distribution.

The mean location is given by

$$\langle c \rangle = \frac{2(\gamma-1)}{(t_{max}^2(\gamma+1))}\left[\int_{s=0}^{t_{max}} s^2 ds + t_{max}(\frac{1}{t_{max}})^{-\gamma}\int_{s=t_{max}}^{\infty} s^{-\gamma+1} ds\right],$$

which is

$$\langle c \rangle = \frac{2(\gamma-1)}{(t_{max}^2(\gamma+1))}\left[[\frac{s^3}{3}]_{s=0}^{t_{max}} + t_{max}(\frac{1}{t_{max}})^{-\gamma}[\frac{s^{-\gamma+2}}{-\gamma+2}]_{s=t_{max}}^{\infty}\right],$$

and, provided $\gamma > 2$, gives

$$\langle c \rangle = \frac{2(\gamma-1)}{(t_{max}^2(\gamma+1))}\left[[\frac{t_{max}^3}{3}] + ((t_{max})^{\gamma+1}[-\frac{t_{max}^{-\gamma+2}}{-\gamma+2}]\right],$$

and finally

$$\langle c \rangle = \frac{2(\gamma-1)t_{max}}{(\gamma+1)}\left[\frac{1}{3} + \frac{1}{\gamma-2}\right] = \frac{2(\gamma-1)t_{max}}{3(\gamma-2)}; \quad \gamma > 2 \tag{0.28}$$

There is little point in computing higher moments because they place even greater constraints on the value of γ (they diverge unless the support for the distribution is a finite interval), and will not apply to our examples for which $2 \le \gamma \le 4.5$, (recall that if the pdf has slope $-\gamma$, the ccdf will have slope $-\beta + 1$ and we are measuring from the ccdf following [12]).

Applying the above estimate for $\langle c \rangle$ to the full Trembl distribution, Fig. 1a of the main text, for which $\gamma = 4.14$ suggests that $\langle c \rangle / t_{max} \approx 1$, where as analysis of the data itself in R gives a ratio of around 1.5 which is reasonable given the nature of the approximation. The actual values from Trembl are given in Table 1.

Table 1: Different measures of average length of proteins in amino acids in the full Trembl 17-03 distribution, Fig. 1a of the main text

	Mean	Median	Mode
Trembl	335	268	219

Homogeneous boxes of beads: CoHSI and Zipf's law

There are some kinds of discrete system for which the heterogeneous model does not apply. Consider the case of homogeneous components. Here, each bead carries a payload such that each box contains *only beads with the same payload, unique to that box*. We represent this by assembling beads of the same colour in the appropriate box, Fig. 7.

Figure 7: The homogeneous case. In each box, all the beads are the same but different boxes contain different types and numbers of beads. This is relevant to the distribution of atomic elements and to the rank ordering of frequency occurrence of words in texts.

We could of course simply set $a_i = 1$ in the heterogeneous case above. However, this immediately causes problems because the asymptotic Hartley-Shannon information content of any box in this case would be $t_i \log 1 = 0$ and is simply degenerate. However, because Hartley-Shannon information is simply the log of the number of ways of arranging the beads of a box, in the absence of an alphabet of choices in each box we can still find a suitable non-degenerate definition as follows.

Suppose we have a unique alphabet of beads $a'_i, i = 1, .., M$ *for the system as a whole*. This is in contrast to the heterogeneous case where the unique alphabet a_i was relevant only to the i^{th} box. Suppose from this system-wide alphabet, we seek to fill the M boxes each with t_i of the a'_i beads such that each box contains only one type. The total population of the M boxes is as before $T = \sum_{i=1}^{M} t_i$. We will renumber them without loss of generality so that $t_1 \leq t_2 \leq .. \leq t_M$.

We proceed as follows. Select any box and then fill it by selecting t_M beads of the same colour. Since we are selecting from M different beads, the probability that we will achieve this selecting at random is $(1/M)^{t_M}$. For the second box, we then have an alphabet available of $M - 1$, so the probability of filling this box with only one colour of the remaining colours is $(1/(M-1))^{t_{M-1}}$ and so on.

The total number of ways N_h this can be done is then given by this probability multiplied by the total number of ways in which T beads can be selected, which is $T!$.

$$N_h = T!\left[(\frac{1}{M})^{t_M} \times (\frac{1}{M-1})^{t_{M-1}} \times .. \times (\frac{1}{1})^{t_1}\right] = T! \prod_{i=1}^{M} (\frac{1}{i})^{t_i} \tag{0.29}$$

Rewriting (0.29) then, the information content of this system is

$$\log N_h = \log T! + \sum_{i=1}^{M} t_i \log\left(\frac{1}{i}\right) = \log T! - \sum_{i=1}^{M} t_i \log i \tag{0.30}$$

The development (0.12)-(0.16) then follows but with $\log i$ replacing $\log a_i$. The end result is the equivalent of (0.16) and amounts to

$$t_i \sim i^{-\eta}, \tag{0.31}$$

where η is some constant.

(0.31) states that if we organise these homogeneous boxes in rank order of contents, (i.e. fullest first), then it is overwhelmingly likely that they will be distributed as a power-law in that *rank*. This is a famous law known as Zipf's law [13]. Zipf's law is empirical although others have produced statistical derivations [14,15]. The above derivation therefore serves as an alternative theoretical justification which places it nicely amongst those distributions which can be explained by the approach taken in this paper.

Appendix B: CoHSI and Implications for Average Component Length and Long Components

Average component length

It has been observed experimentally on several occasions [6,16,17] that proteins appear to preserve their average length across aggregations within relatively tight bounds. The sharply unimodal peak of Figs. 1a-b of the main text as predicted by the theoretical development in this paper suggests that we should not be surprised at this. Indeed at all scales and ensembles the estimates of average protein length will be highly conserved within collections as a result, even though the position of the peak may move a little as shown in Figs 6a, 6b.

In some aggregations, the degree to which the average length is preserved is quite remarkable, for example in Bacteria (Fig. 8a), whilst in Eukarya, there is evidence of some fine structure [6] which invites further analysis Fig. 8b.

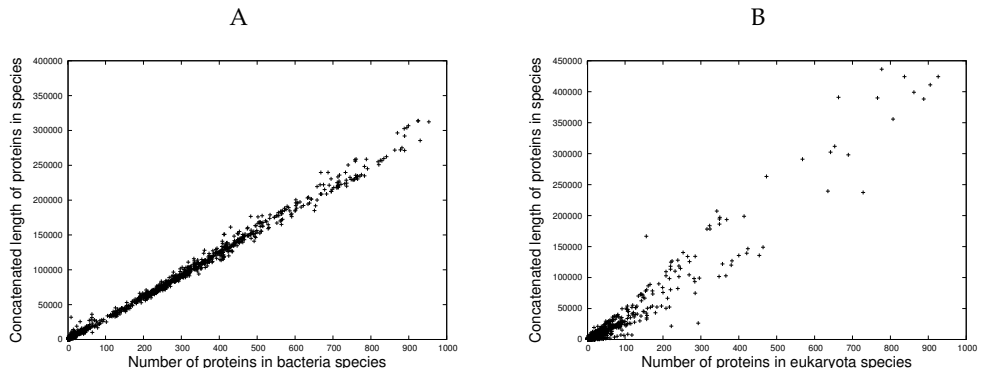

Figure 8: A plot of the total concatenated length of proteins against the total number of proteins for each species in (A) Bacteria and (B) Eukaryota. Each data point corresponds to a species. The gradient of the linearity evident in both plots effectively defines the average protein length for that collection, from [6].

We also note that preservation of the average component length has also been reported for software [4].

Measuring average protein length

The skewed nature of the distribution of Figs. 1a-b suggests that the use of the mean as a measure of average length alone may be misleading and should be accompanied by other more robust measures such as the median and mode. Table 2 demonstrates this by calculating them for the three domains of Archaea, Bacteria, Eukaryota, along with Viruses, as shown in Figs. 4a-d of the main text. As expected, the more robust measure of median is less affected by the skew and the medians are therefore considerably less spread out than the means. This is particularly true of viruses which although they have an anomalously large mean in comparison, their median is much closer aligned with those of Archaea and Bacteria. The modes are subject to considerable noise.

Table 2 shows that the mean is around $1-3$ times the modal value.

Table 2: Different Measures of average length of proteins in the domains of life and viruses

Domain	Mean	Median	Mode
Archaea	287	246	130
Bacteria	312	272	156
Eukaryota	435	350	379
Viruses	451	289	252

Long components

One of the most important features of power-laws compared with any kind of exponential distribution such as the normal distribution is that "events that are effectively 'impossible' (negligible probability under an exponential distribution) become practically commonplace under a power-law distribution." [18]. The emphatic power-law in both the protein lengths and in software function lengths inevitably leads to large ratios when comparing the longest components with the average. For example, proteins of around $36,000$ amino acids have been found and this is $100\times$ the average.

In terms of the theory developed here, there is no need for any biological reason for very long proteins - they exist simply because of the naturally emerging power-law resulting from consideration of Information-conserving ergodic systems.

We note that precisely the same thing has been observed in software [19].

Appendix C: CoHSI and Token Alphabets

The definition of alphabets, i.e., unique sets of tokens from which choices can be made, poses interesting questions. First of all, we must point out that there is generally no obvious definitive unique alphabet for any system. Alphabets are partly subjective and partly objective because at their heart, they are about how humans categorise systems. Take a simple example of a normally sighted person and a colour blind person both counting the number of differently coloured beads in a collection. Barring counting errors, they will both find the same number of beads in total, however, *they will not necessarily agree on the number of beads of each colour.* In particular, red-green confusion is likely[2].

How does this affect the theory we describe here ? It might be thought that by linearly increasing the *size* of the alphabet, the distribution of the two alphabets are themselves linearly related, i.e. $alphabet1_i = constant \times alphabet2_i$ However, this turns out to be *not* the case and to understand what is happening, we must return to the duality of the asymptotic behaviour of length and alphabet distributions described in the main text, which we repeat here,

$$p_i \equiv \frac{t_i}{T} = \frac{a_i^{-\beta}}{Q(\beta)} \tag{0.32}$$

and its algebraic dual given by

$$q_i \equiv \frac{a_i}{A} = \frac{t_i^{-1/\beta}}{\sum_{i=1}^{M} t_i^{-1/\beta}} \tag{0.33}$$

Since our normally-sighted person and our colour-blind person will count the same numbers but with different alphabets, we can say that for the normally sighted person,

[2] https://en.wikipedia.org/wiki/Color_blindness

$$p_i \equiv \frac{t_i}{T} = \frac{(a_i')^{-\beta'}}{Q(\beta')} \tag{0.34}$$

and for our colour blind person

$$p_i \equiv \frac{t_i}{T} = \frac{(a_i'')^{-\beta''}}{Q(\beta'')} \tag{0.35}$$

where a_i', a_i'' are the two unique alphabets they use and β', β'' their slopes. Since the lengths are unchanged, we can see straight away from (0.34) and (0.35), that the two unique alphabets will themselves be power-law related asymptotically

$$(a_i')^{-\beta'} \sim (a_i'')^{-\beta''} \Rightarrow a_i' \sim (a_i'')^{-\beta'''}, \tag{0.36}$$

where $\beta''' = -\beta''/\beta'$. This leads us to predict a general rule

In any consistent categorisation of the same system with different unique alphabets, the distributions of the unique alphabets will also be related by a power-law.

Consider an example from the world of music.

Music alphabets

Music is also a system of discrete components in the sense described here in this paper. In recent years, discrete formats representing the notes and structure of a musical composition have appeared, for example MusicXML as referenced in the main text.

If we consider the 88 notes of a full-scale piano as defining the possible notes in the equal-tempered scale used in the vast majority of published music, then we have a candidate unique alphabet a_i' of 88 *(no-duration alphabet)*. However, we can subdivide this alphabet quite naturally and consistently into notes *and* duration. The standard durations are divided into fractions of a whole note as breve (2), semi-breve (1), minim (1/2), crotchet (1/4), quaver (1/8), semiquaver (1/16) and demisemiquaver (1/32). There are others defined off either end of this list but they are obviously rare as there were no occurrences in the body of music studied here. This gives seven flavours of each note and expands the unique alphabet considerably to $88 \times 7 = 616$ items, *(duration alphabet)*.

Figs. 9a shows the distribution of the two alphabets *no-duration* and *duration*, measured on the same body of music. As expected from (0.34) and (0.35) they both exhibit power-law behaviour.

For the *no-duration* alphabet R reports that the associated p-value matching the power-law tail linearity in the ccdf of Fig. 9a is $< (2.2) \times e^{-16}$ over the range $40.0 - 100.0$, with an adjusted R-squared value of 0.8482. The slope is -4.20 ± 0.26. For the better-populated *duration* alphabet R reports that the associated p-value matching the power-law tail linearity in the ccdf of Fig. 9a is $< (2.2) \times e^{-16}$ over the range $40.0 - 1000.0$, with an adjusted R-squared value of 0.9459. The slope is -1.50 ± 0.03.

Moreover in Fig. 9b which compares the two alphabets directly on a $\log - \log$ scale, the predicted power-law relationship of (0.36) is clearly visible. R reports that the associated p-value matching the power-law tail linearity in the ccdf of Fig. 9b is $< (2.2) \times e^{-16}$ over the range $10.0 - 500.0$, with an adjusted R-squared value of 0.9754. The slope is 1.60 ± 0.01, also consistent with (0.36).

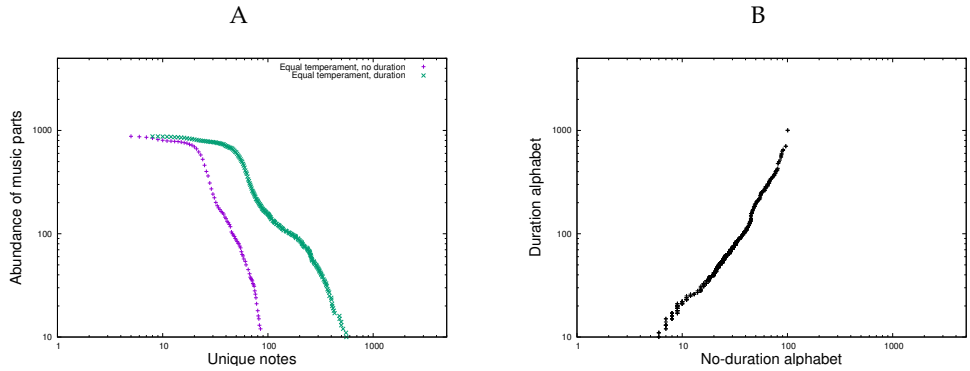

Figure 9: (A) The log − log ccdf of the duration and no-duration alphabets measured on the same body of music used in this study and, (B) a comparison of the two alphabets as log − log showing their clear linear power-law relationship.

Considering that the music dataset is by far the smallest heterogeneous system analysed in this paper, these are still satisfyingly positive results. We believe that this throws some light (but does not necessarily explain) why the predicted reciprocal relationship between the power-law slope of the unique alphabet distribution and that of the length distribution [6] is not adhered to closely in our data. *There are a potentially infinite number of alphabets related themselves by power-laws, but only one length distribution.* The future greater availability of MusicXML data will help clarify this.

Protein alphabets

For proteins, with growing sophistication we are able to recognise not just the 22 amino acids transcribed directly from DNA but also the increasingly large number of known post-translational modifications (PTM) which dramatically extend and continue to extend the size of the unique amino acid alphabet that allows us to categorise proteins. Will this process of discovery stop ? We argue that it cannot as it is intimately linked to the total number of proteins known, and while this continues to grow apace, so will the known number of PTM amino acids.

We can gain insight into the growth of the unique protein alphabet by studying collections, such as the SwissProt database, over different revisions [20–22] as it incorporates PTM information from the Selene project [23].

Fig. 10 shows the frequencies of the unique amino acid alphabet recorded in the proteins of SwissProt release 13-11 and SwissProt release 15-07 as ccdfs in log − log form. Although the maximum unique alphabet for amino acids is 22 for those decoded directly from DNA, we should note that finding a protein with all 22 would be unlikely as both the 21^{st} and 22^{nd} amino acids, selenocysteine and pyrrolysine, are rare in proteins. Pyrrolysine is found in methanogenic archaea and bacteria and is encoded by a re-purposed stop codon (UAG), requiring the action of additional gene products to accomplish its incorporation [24]. Thus it is not easy to annotate pyrrolysine from the gene sequence alone, and direct chemical analysis of proteins would be more informative.

Selenocysteine is found in all domains of life, but the selenoproteome is small [25] and an additional concern is misannotation in the databases, because a stop codon (UGA) is re-purposed from “halt translation” to “incorporate seleocysteine” by additional sequences downstream of the gene as well as other trans-acting factors [26].

As a result, any unique amino acid count beyond 21 must contain post-translationally modified amino acids and we note the following:

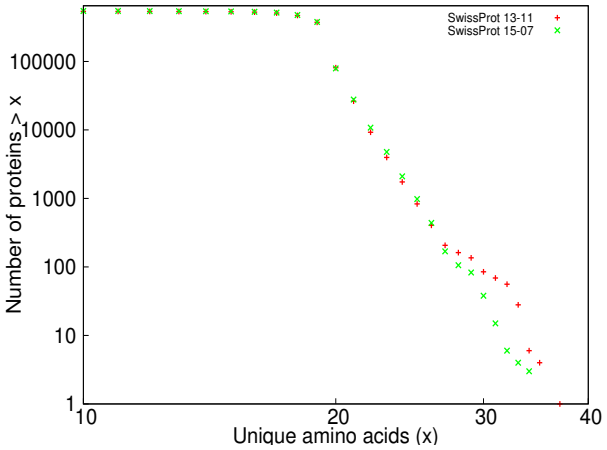

Figure 10: The classic linear signature of a power-law distributions of unique amino acid alphabets for SwissProt 13-11 and the more up to date SwissProt 15-07 on a $\log - \log$ ccdf.

- Almost the whole of the tail of Fig. 10 consists of proteins in which there must be post-translationally modified amino acids, effectively doubling the unique alphabet derived directly from DNA.
- The increase in numbers between 20 and 26 unique amino acids can be seen by the slightly displaced points upwards in the SwissProt 15-07 dataset compared with the SwissProt 13-11 dataset.
- The remainder of the tail significantly straightens with the more comprehensively annotated SwissProt 15-07 presumably due to reduction in noise with increasing numbers.

The linearity in each tail strongly supports power-law behaviour even though the range is less than 1 decade because the slope is so steep arising from the current paucity of the unique alphabet. For SwissProt 13-11 of Fig. 10, R lm() reports that the associated p-value matching the power-law tail linearity is $8.124 \times e^{-13}$ **over the range** $19 - 35$**, with an adjusted R-squared value of** 0.9698**. The slope is** -15.91 ± 0.56**. For SwissProt 15-07 of Fig. 10, R lm() reports that the associated p-value matching the power-law tail linearity is** $< 2.2 \times e^{-16}$ **over the range** $19 - 33$**, with an adjusted R-squared value of** 0.9968**. The slope is** -18.56 ± 0.19**.**

We note in passing that although the distribution of unique amino acids has a small power-law tail, the distribution is anything but uniform as we can see in SwissProt 13-11 by considering Fig. 11a which plots the logarithmic frequency of proteins against their unique amino acid alphabet. Whatever this distribution is, it is certainly not uniform, although we know the tail from $19 - 35$ amino acids is an accurate power-law from the analysis of the data shown in Fig. 10.

In contrast, Fig. 11b plots the occurrence rate of each unique amino acid including post-translational modification across the entire SwissProt 13-11 distribution, of which there are more than 800 recorded by the Selene project. In other words it shows in how many proteins each amino acid appears, organised in rank order. This matches the homogeneous model discussed in Appendix A p. 13, and a power-law in the tail is evident as expected.

We note in passing a possible intriguing relationship between the overhang in Fig. 11b from around 10 to 30 on the x-axis and the contemporary question of PTM undercounting [27], although we will not pursue this further here.

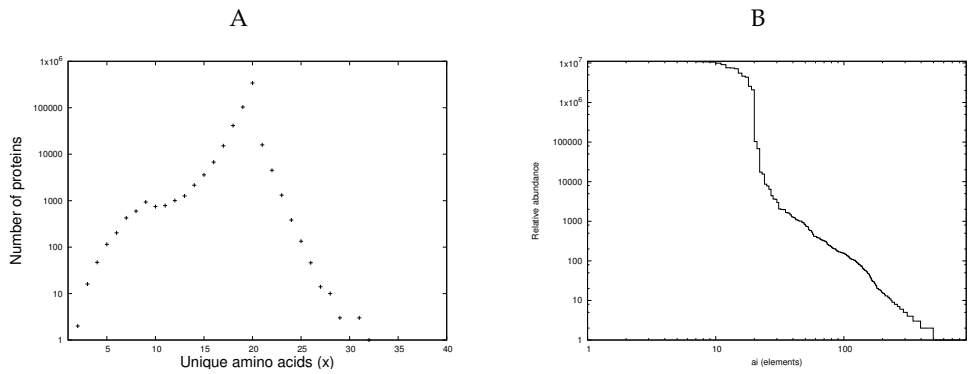

Figure 11: (A) The frequency of proteins plotted against the unique amino acid count for SwissProt 13-11 on a $\log -linear$ plot and, (B) the frequency at which each amino acid occurs including PTM amino acids plotted in rank order on a $\log - \log$ ccdf (B).

R lm() reports that the associated p-value matching the power-law tail linearity in the ccdf of Fig. 11b is $< (2.2) \times e^{-16}$ over the range $22.0 - 800.0$, with an adjusted R-squared value of 0.9778. The slope is -2.63 ± 0.31. This too is an emphatic result.

Appendix D: Power-laws, Statistical Rigour and Rules of Thumb

Power-laws are ubiquitous in nature and are generated by a number of mechanisms, [12]. In essence, power-law behaviour can be represented by the pdf (probability density function) p(s) of entities of size s appearing in some process, given by a relationship like

$$p(s) = \frac{k}{s^b} \tag{0.37}$$

where $k, b(>1)$ are constants. On a $\log p - \log s$ scale the pdf is a straight line with negative slope $-b$. It can easily be verified that the equivalent cdf (cumulative density function) $c'(s)$ derived by integrating (0.37) also obeys a power-law $\sim s^{-b+1}$, (for $b \neq 1$). The classic linear signature of a power-law tail $c(s)$ in a ccdf (complementary cumulative distribution function) is usually shown as in Fig. 12 which displays $c(s) = 1 - c'(s)$.

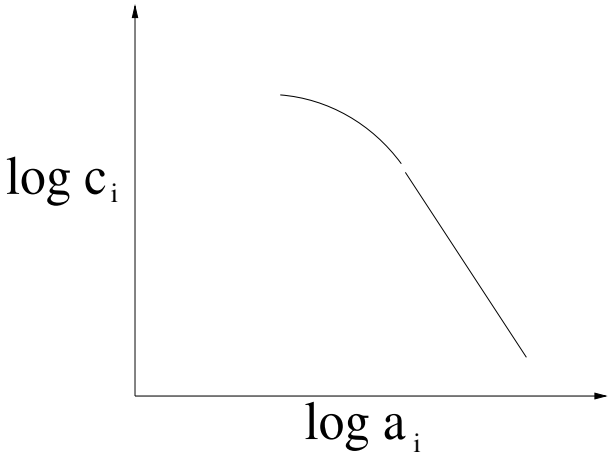

Figure 12: The classic linear signature of a power-law in the tail of a $\log - \log$ ccdf.

For noisy data, the ccdf form is used most often because of its fundamental property of reducing noise present in the pdf, as noted by [12]. This effect is because the ccdf is obtained by integration. This reduces noise inherent in the pdf preserving any power-law behaviour while allowing any linearity to be measured more accurately. The benefit of this can be clearly seen in data extracted from software systems as shown in Figs. 13a (the pdf) and 13b (the corresponding ccdf). The effect is even more pronounced in the considerably more noisy protein data, (due to experimental error).

Whilst on the subject of significance, a rule of thumb often used to determine the existence of a power-law is that it should appear over two or more decades in the x-axis of the ccdf. This is useful only as a rule of thumb when the slope is not too steep. Since the scale of the y-axis on the $\log - \log$ ccdf is effectively the scale of the x-axis times the slope of the power-law, then a steep slope would require a large scale of y-axis frequency measurements to provide the rule of thumb of 2-3 decades in the x-axis. For example, a slope of around 3 would require y-axis frequency measurement over some 6-9 decades to give the rule of thumb of 2-3 decades in the x-axis, which is reasonable. On the other hand, if the slope is 10, y-axis frequency measurement over some unreasonably large 20-30 decades would be required to give the rule of thumb of 2-3 decades in the x-axis.

This is an important point for the protein studies considered here wherein we are investigating the predicted power-law in unique alphabet. Here, the x-axis is the unique alphabet of amino acids. The size of this alphabet is small *at the current state of knowledge*, leading to a steep power-law slope. In such situations, we fall back on normal procedures of statistical inference

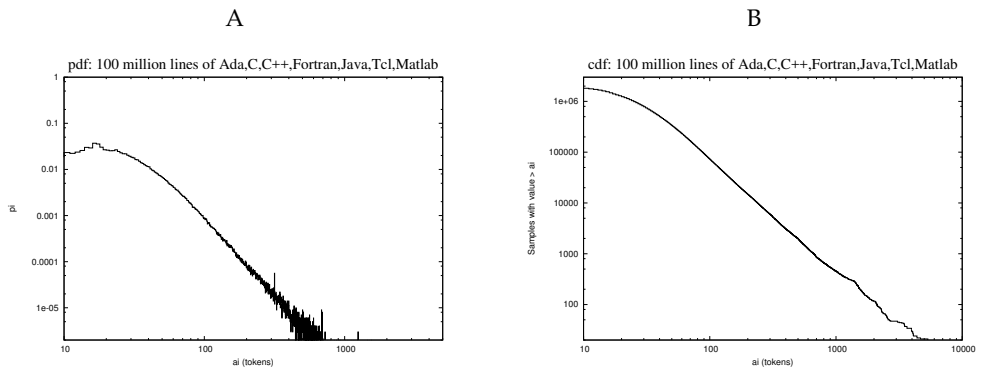

Figure 13: The pdf (A) and the ccdf (B) of the length distributions of the same large population of software.

to replace subjective belief with objective perception and therefore all that is required is that there is statistically significant linearity in the tail of the distribution of the $\log - \log$ ccdf for the number of measurement points used. A rule of thumb guides but does not replace normal statistical inference whereby a result is either significant at some level or it is not for a given model and data.

This effect can be seen in Fig. 14, a $\log - \log$ ccdf of the occurrence rate in the size of the unique alphabet in SwissProt version 13-11 [20], which is merged with the Selene post-translational modification data [23,28]. These represent amongst the best annotated protein data including the rapidly growing field of post-translational modification (PTM), a process whereby nature alters some of the amino acids by covalent processes such as glycosylation, phosphorylation, methylation, acylation, etc., thereby extending the unique alphabet beyond the 22 amino acids directly coded from DNA [29–33].

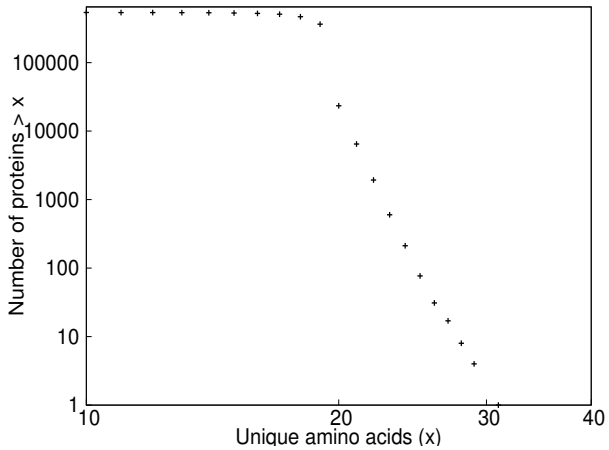

Figure 14: The highly linear tail of the occurrence frequency of unique alphabet sizes in the SwissProt 13-11 protein distribution merged with the Selene 2013 post-translational modification annotation, extending the range of the natural unique alphabet of 22 amino acids directly coded from DNA to just over 30 in this dataset.

As can be seen, the tail of SwissProt 13-11 in Fig. 14, covers only a range up to just over 30 even though there are thousands of PTM known or predicted by existing research. As a result there are only nine data points for unique alphabets of size greater than 20, although each point is an aggregate of a large number of observations.

An R lm() analysis on this tail reports that the associated p-value matching the power-law tail linearity in the ccdf of Fig. 9a is $6.576 \times e^{-12}$ over the range $21.0 - 30.0$, with an adjusted R-squared value of 0.9951. The slope is -22.9 ± 0.2.

This is a statistically emphatic result for the existence of a power-law, even though the size of the slope is steep because the x-axis is restricted. It would be perfectly possible of course that another distribution other than power-law might fit the data better, but we are not simply data-fitting here. Instead, we are testing a theory which predicts a power-law in the tail of the distribution to see if the observations are consistent with that prediction. If we are able to prove that they are not, then the theory is deficient. So far, we have found no evidence to reject our theory. Note that later versions of SwissProt with Selene annotations increase the PTM alphabet, Appendix C p. 18, thereby decreasing this slope.

Power-law behaviour has been studied in a wide variety of environments starting with the pioneering work of [13] (linguistics) and followed by [34] (economic systems) and the excellent reviews by [35] and [12]. In software systems significant activity, much of it recent, [36], [11], [37], [38], [39], [40], [41], [42], [19] and [43] has addressed power-law behaviour in various contexts.

To give some idea of the scope of these, Mitzenmacher [11] considers the distributions of file sizes in general filing systems and observed that such file sizes were typically distributed with a lognormal body and a Pareto (i.e. power-law) tail. Gorshenev and Pis'mak [39] studied the version control records of a number of open source systems with particular reference to the number of lines added and deleted at each revision cycle. Louridas et. al. [19] show evidence that power laws appear in software at the class and function level and that distributions with long, fat tails in software are much more pervasive than previously established.

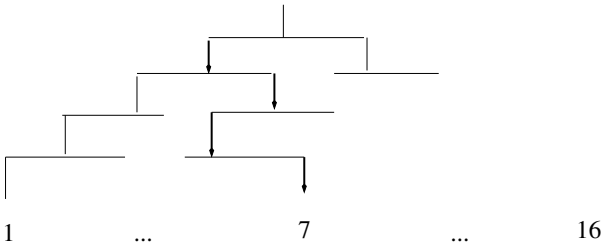

Figure 15: A binary tree. Each level proceeding down can either go left or right. There are four levels leading down to one of 2^4 = 16 possibilities. Only four choices are needed to reach any of the possibilities. We note that $\log_2(16) = 4$. Here the number 7 has been singled out by the four choices left, right, left, right as the tree is descended.

Appendix E: Hartley-Shannon Information, parsimony and token-agnosticism

Information theory has its roots in the work of Hartley [5] who showed that a message of N signs (i.e. tokens) chosen from an *alphabet* or code book of S signs has S^N possibilities and that the *quantity of information* is most reasonably defined as the *logarithm* of the number of possibilities or choices $\log S^N = N \log S$. To gain insight into the reason why the logarithm makes sense, consider Fig. 15. The number of choices necessary to reach any of the 16 possible targets is the number of levels which is $\log_2$(number of possibilities). The base of the logarithm is not important here.

Information theory was developed substantially by the pioneering work of Shannon [44], [45] and many researchers since but for the greatest generality, we have remained with Hartley's original clear vision and most importantly its token-agnosticism. We re-iterate that is important not to conflate information content with functionality or meaning and Cherry [46] specifically cautions against this noting that the concept of information based on alphabets as extended by Shannon and Wiener amongst others, *relates only to the symbols themselves* and not their *meaning*. Indeed, Hartley in his original work, defined *information* as the successive selection of signs, rejecting all meaning as a mere subjective factor. In the sense used here therefore, Conservation of H-S Information will be synonymous with Conservation of Choice, not meaning. In spite of its simplicity, this turns out to be enough to predict the important system properties detailed in this paper. In other words, those properties depend only on the alphabet and not on what combining tokens of the alphabet might mean in any human sense.

We believe CoHSI therefore represents the most parsimonious theory capable of explaining all the observed features of the numerous disparate datasets analysed in this paper.

References

1. Glazer A, Wark J. 2001 *Statistical Mechanics. A survival guide*. OUP.
2. Sommerfeld A. 1956 *Thermodynamics and Statistical Mechanics*. Academic Press.
3. Frank SA. 2009 The common patterns of nature. *Journal of Evolutionary Biology* **22**, 1563–1585. 10.1111/j.1420-9101.2009.01775.x.
4. Hatton L. 2014 Conservation of Information: Software's Hidden Clockwork. *IEEE Transactions on Software Engineering* **40**, 450–460. 10.1109/TSE.2014.2316158.
5. Hartley R. 1928 Transmission of Information. *Bell System Tech. Journal* **7**, 535.
6. Hatton L, Warr G. 2015 Protein Structure and Evolution: Are They Constrained Globally by a Principle Derived from Information Theory ?. *PLOS ONE*. doi:10.1371/journal.pone.0125663.
7. Ramanujan S. 1988 *The lost notebook and other unpublished papers*. Springer-Verlag. ISBN 978-3-540-18726-4.
8. Tsallis C. 1988a Possible Generalizations of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **52**, 479–87.
9. Tsallis C. 1988b Nonextensive statistics: Theoretical, experimental and computational evidences and connections. *Braz. J. Phys.* **29**, 1–35.

10. Montroll E, Schlesinger M. 1982 On 1/f noise and other distributions with long tails. *Proc. Nat. Acad. Sci. USA* **79**, 3380–3.
11. Mitzenmacher M. 2002 Dynamic Models for File Sizes and Double Pareto Distributions. *Internet Mathematics* **1**, 305–333.
12. Newman MEJ. 2006 Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46**, 323–351.
13. Zipf G. 1935 *Psycho-Biology of Languages*. Houghton-Miflin.
14. Simon H. 1955 On a Class of Skew Distribution Functions. *Biometrika* **42**, 425–440.
15. Li W. 1992 Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Transactions on Information Theory* **38**, 1842–1845. DOI: 10.1109/18.165464.
16. Wang D, Hsieh M, Li W. 2005 A general tendency for conservation of protein length across eukaryotic kingdom. *Molecular Biology and Evolution* **22**, 142–147. 10.1093/molbev/msh263.
17. Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo Z. June 2006 Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. *Molecular Biology and Evolution* **23**, 1107–1108. 10.1093/molbev/msk019.
18. Clauset A. 2011 Inference, Models and Simulation for Complex Systems. Lectures. http://tuvalu.santafe.edu/ãaronc/courses/7000/csci7000-001_2011lL2.pdf, accessed 24-Jun-2017.
19. P. Louridas P, Spinellis D, Vlachos V. 2008 Power Laws in Software. *ACM Trans. Softw. Eng. Methodol.* **18**, 2:1–2:26. 10.1145/1391984.1391986.
20. SwissProt. 2013 The SwissProt release, 13-11. SwissProt http://www.uniprot.org/.
21. SwissProt. 2015 The SwissProt release, 15-07. SwissProt http://www.uniprot.org/.
22. SwissProt. 2017 The SwissProt release, 17-03. SwissProt http://www.uniprot.org/.
23. Selene Project T. 2013 . http://selene.princeton.edu/PTMCuration/.
24. Quitterer F, List A, Beck P, Backer A, Groll M. 2012 Biosynthesis of the 22nd genetically encoded amino acid pyrrolysine: structure and reaction mechanism of PylC at 1.5ÃĚ resolution. *J. Mol. Biol* **424**, 270–82. DOI: 10.1016/j.jmb.2012.09.007.
25. Reeves M, Hoffmann P. 2009 The human selenoproteome: recent insights into functions and regulation. *Cell Mol Life Sci.* **66**, 2457–78. DOI: 10.1007/s00018-009-0032-4.
26. Mehta A, Rebsch C, Kinzy S, Fletcher J, Copeland P. 2004 Efficiency of mammalian selenocysteine incorporation. *J Biol Chem* **279**, 37852–9. DOI: 10.1074/jbc.M404639200.
27. Thaysen-Andersen M, Packer N. 2014 Advances in LC-MS/MS- based glycoproteomics: Getting closer to system-wide site-specific mapping of the N- and O-glycoproteome. *Biochim Biophys Acta* **1844**, 1437–1452.
28. SwissProt. 2014 Controlled vocabulary of posttranslational modifications PTM. http://www.uniprot.org/docs/ptmlist.
29. Apweiler R, Hermjakob H, Sharon N. 1999 On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* **1473**, 4–8.
30. Khoury G, Baliban R, Floudas C. 2011 Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep. 1* **1**. 10.1038/srep00090.
31. Zafar S, Nasir A, Bokhari H. 2011 Computational analysis reveals abundance of potential glycoproteins in Archaea, Bacteria and Eukarya. *Bioinformation* **6**, 352–355.
32. Prabakaran S, Lippens G, Steen H, Gunawardena J. 2012 Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding. *WIREs Syst Biol Med* **4**, 565–583. 10.1002/wsbm.1185.
33. Campbell M, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita K, Lisacek F, Packer N. 2014 UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res*. 10.1093/nar/gkt1128.
34. Rawlings P, Reguera D, Reiss H. 2004 Entropic basis of the Pareto law. *Physica A* **343**, 643–652.
35. Mitzenmacher M. 2003 A brief history of generative models for power-law and lognormal distributions. *Internet Mathematics* **1**, 226–251.
36. Clark D, Green C. 1977 An empirical study of list structures in Lisp. *Communications of the ACM* **20**, 78–87.
37. Myers CR. 2003 Software systems as complex networks: Structure, function and evolvability of software collaboration graphs. *Physical Review E* **68**.
38. Challet D, Lombardoni A. 2004 Bug propagation and debugging in asymmetric software structures. *Physical Review E* **70**.

39. Gorshenev A, Pis'mak YM. 2004 Punctuated equilibrium in software evolution. *Physical Review E* **70**, 067103–1,4.
40. Potanin A, Noble J, Frean M, Biddle R. 2005 Scale-free geometry in OO programs. *Comm. ACM.* **48**, 99–103.
41. Baxter G, Frean M, Noble J, Rickerby M, Smith H, Visser M, Melton H, Tempero E. 2006 Understanding the shape of Java software. *OOPSLA '06*. http://doi.acm.org/10.1145/1167473.1167507.
42. Concas G, Marchesi M, Pinna S, N.Serra. 2007 Power-Laws in a Large Object-Oriented Software System. *IEEE Transactions on Software Engineering* **33**, 687–708.
43. Hatton L. 2009 Power-Law distributions of component sizes in general software systems. *IEEE Transactions on Software Engineering* **35**, 566–572. http://doi.ieeecomputersociety.org/10.1109/TSE.2008.105.
44. Shannon C. 1948 A mathematical theory of communication. *Bell System Tech. Journal* **27**, 379–423.
45. Shannon C. 1949 Communication in the Presence of Noise. *Proc. I. R. E.* **37**, 10.
46. Cherry C. 1963 *On Human Communication*. John Wiley Science Editions. Library of Congress 56-9820.