RESEARCH ARTICLE

New concept for quantification of similarity relates entropy and energy of objects: First and Second Law entangled, group behavior of micro black holes expected

Petr Zimak, Silvia Terenzi, Peter Strazewski[*]

Abstract

When the free energy of similar but distinct molecule-sized objects is plotted against the temperature at which their energy and entropy contributions cancel, a highly significant linear dependence results from which the degree of similarity between the distinctly different members within the group of objects can be quantified and a relationship between energy and entropy is derived. This energy-entropy relationship entirely reflects the mathematical structure of thermodynamic equations, is in this sense fundamental and therefore does probably not dependent on material nor scale. The energy-entropy relationship is likely to be of general interest in molecular biology, population biology, synthetic biology, biophysics, chemical thermodynamics, systems chemistry and physics, most notably in particle physics and cosmology. In physics we predict a consistent and perhaps testable way of classifying micro black holes, to be generated in future Large Hadron Collider experiments, by their gravitational energy and area entropy.

Introduction

The larger the physical scale is, the less frequently the term 'energy' and the more frequently the term 'entropy' is used in physics discussions. Energy, in the sense of 'bound' or 'inner' energy, is an entity that is usually measured experimentally in some more or less direct way. Entropy is an entity impossible to measure directly; it can only be determined either in conjunction with measured energy and another measured experimental parameter, free energy for instance, or it is calculated or counted using statistical mechanics or some other theory on the degeneracy of microstates. Since, owing to their distance from the observer, very large-scale physical objects are difficult to measure directly, the preferential use of entropy and the Second Law of thermodynamics is not astonishing in cosmology, neither is the preferential use of energy in quantum physics, in particular, strict energy conservation as expressed through the First Law of

thermodynamics. Of course both laws apply *a priori* to all scales and physics, and of course the above statements are not based on statistical analyses or other objective grounds but on the subjective impression of the author to whom correspondence should be addressed.

In this article we present very briefly the results of a comprehensive analysis of published experimental thermodynamic data on the unfolding of many hundreds of proteins and nucleic acids, on molecular associations in host-guest complexes, on the stability of ab initio (quantum mechanically) calculated water clusters and the semi-empirically (force field) calculated formation thermodynamics of small organic molecules from their elements. We then mainly discuss the consequences when i) these numerical results are first grouped into families that distinguish ensembles of evidently similar objects, ii) the grouped results are correlated in a specific two-dimensional projection of a five-dimensional parameter space and, ultimately, iii) the results are detached from the molecular scale.

The discussion begins with deriving an equation that relates energy changes to entropy changes of the same

* Correspondence: strazewski@univ-lyon1.fr
Laboratoire de Synthèse de Biomolécules, Institut de Chimie et Biochimie Moléculaires et Supramoléculaires (CNRS UMR 5246), Université Claude Bernard Lyon 1, Université de Lyon, 43 bvd du 11 novembre 1918, F - 69622, Villeurbanne, France

objects without usage of additional empirical parameters or functions that are not explained from the fundamentals. The only new 'entity' or 'information' is the fact that the objects are grouped into families of obviously similar characteristics. Protein mutants and nucleic acid variants are macromolecules that usually differ only very little in overall shape and folding potential - only one or two in dozens or hundreds of 'chain links' are different within the same group - but may differ rather heavily in measured energy and entropy of folding. It is known since 1970 that in many very different chemical and biological systems large entropy and energy contributions compensate one another, to give small resulting free energy changes, that is, small net effects. We do not discuss this here - our studies on the compensation effect and statistical significance of the utilised linear regressions are described in full detail to be published elsewhere - but rather focus on the consequences of the results. Once energy and entropy changes are fundamentally linked to one another, the laws that on the one hand restrict in isolated systems average net energy changes to zero and on the other hand confine spontaneous net entropy changes to zero or more but not less, thus, condemn entropy to maximise over time, may become fundamentally linked as well. If our analysis on the thermodynamics of medium-sized objects, which can either be described by quantum physics or by classical physics, were generalizable to all scales, we were to conclude the following.

The First and Second Law of thermodynamics describe isolated multicomponent systems in the observable universe as objects that conserve their energy due to their very isolation *and* that spontaneously maximise their entropy over time. For the latter to be true, the objects' size must be sufficiently large for fully reversible changes, that is, exactly reversed changes in their microstates, to become too improbable to occur within their lifetime. Additionally, an *isolated ensemble of similar objects* in the same universe will spontaneously maximise its overall entropy over time *in a way (at a rate)* that reflects its overall energy *and* identity, thus, its compositional and structural characteristics that define it as an ensemble of similar objects. If the physical isolation of the ensemble confines its overall average energy changes to zero, the way (rate) of maximizing entropy can only change when the degree of similarity within the ensemble of objects changes as well. We conclude that, given a constant (accessed) overall volume of an ensemble, the higher the degree of similarity is among its objects the slower is their rate of spontaneous entropy maximization and the closer to maximum entropy they are. Hence, it seems as if the rate of maximizing overall entropy of an ensemble of objects were

related to the similarity of what characterises the individual objects within the ensemble.

Here we present a statistical means of quantifying the degree of similarity, namely, through the linear regression coefficient obtained from the correlation of the difference with the ratio of two object characterizing parameters (energy U and entropy S) that both depend on one independent variable (absolute temperature T). We depict, using experimental numerical values, 3D projections of the 5D parameter space $\{U; S; T; U - T{\cdot}S; U/S\}_{pV}$ (at constant pressure and volume pV).

Experimental

The vast majority of the primary data are experimental and about one third of those originate from differential scanning calorimetric experiments where both the energy change under constant pressure, i.e., the enthalpy change ΔH, and the position of thermodynamic equilibrium between two macroscopic states, i.e., the free enthalpy change ΔG (Gibbs free energy), are derived from equation 1. The measured heat capacity C_p (at constant pressure) is a function of temperature T within a T-range needed to observe both major macroscopic states (termed 'folded' and 'unfolded') in virtually quantitative abundance. Enthalpy changes in a system open to atmospheric pressure, $\Delta H = H_{\text{macrostate 1}} - H_{\text{macrostate 2}}$, and energy U in a closed system are linked through $U = H - p{\cdot}V$. Likewise, the Gibbs free energy difference $\Delta G = G_{\text{macrostate 1}} - G_{\text{macrostate 2}}$ is a measure for the driving force towards macroscopic stasis under constant pressure, and free energy is linked through $F = G - p{\cdot}V$. The corresponding change in entropy ΔS of the system is usually calculated from $\Delta G = \Delta H - T{\cdot}\Delta S$ (or $\Delta F = \Delta U - T{\cdot}\Delta S$) rather than directly from equation 1.

$$C_p = dH/dT = T{\cdot}dS/dT = -T{\cdot}(d^2G/dT^2) \qquad (1)$$

Another definition of heat capacity is the mean squared fluctuation in energy scaled by kT^2, or the mean squared fluctuation in entropy scaled by k (the Boltzmann constant), as shown in equation 2 [1].

$$C_p = \langle \delta H^2 \rangle / kT^2 = \langle \delta S^2 \rangle / k \qquad (2)$$

The difference in specific heat capacity between both major macroscopic states is directly measured from $\Delta C_p = C_p(T_{\text{100\% unfolded}}) - C_p(T_{\text{100\% folded}})$; Δ always refers to the difference between two distinct macroscopic states. Both $C_p(100\%$ unfolded$)$ and $C_p(100\%$ folded$)$ are assumed to exert the same T-dependence, hence $\partial \Delta C_p/\partial T = 0$, i.e. $\Delta C_p \approx$ const.

The other two thirds of experimental data originate from so-called van't Hoff experiments in which, instead

of C_p, equilibrium constant K = (fraction macrostate 1)/(fraction macrostate 2) = $\exp[-\Delta G/RT]$ (R = 1.9872 cal mol^{-1} K^{-1}) is measured within an appropriate range of T or other parameter capable of completely shifting the thermodynamic equilibrium from one macroscopic state to another. For thermally induced macrostate changes the accompanying energy and entropy changes are elucidated from fitting the experimental data to equation 3:

$$R \cdot \ln K = -\Delta H/T + \Delta S = -\Delta G/T \qquad (3)$$

In the vast majority of published van't Hoff experiments heat capacity changes are ignored altogether: $\Delta C_p \approx 0$. This approximation is justified by the usually observed linear relationship for $\ln K$ versus $1/T$. In both kinds of experiments, calorimetric and van't Hoff, any true T-dependence of ΔC_p may be neglected when compared to the one of $\Delta G = \Delta H - T \cdot \Delta S$ (or of $\Delta F = \Delta U - T \cdot \Delta S$) over the measured T-range. In summary, classical thermodynamics provides us with equations 4 and 5 in the fundamental, most general case $\Delta C_p = f(T)$ [2]. Equations 6 and 7 result from the 'calorimetric neglection' of the T-dependence of ΔC_p. After a 'van't Hoff neglection' of ΔC_p, ΔH and ΔS become constants with respect to T.

$$\Delta H_T = \Delta H_{T_{ref}} + \int_{T_{ref}}^{T} \Delta C_p(T) \mathrm{d}T \qquad (4)$$

$$\Delta S_T = \Delta S_{T_{ref}} + \int_{T_{ref}}^{T} \frac{\Delta C_p(T)}{T} \mathrm{d}T \qquad (5)$$

$$\Delta H_T = \Delta H_{T_{ref}} + \Delta C_p \cdot (T - T_{ref}) \qquad (6)$$

$$\Delta S_T = \Delta S_{T_{ref}} + \Delta C_p \cdot \ln(T/T_{ref}) \qquad (7)$$

Procedure

We extracted from the literature 1555 experimental datasets $\{\Delta C_p; \Delta H_{T_{ref}}; \Delta S_{T_{ref}}\}$ on the thermal and non-thermal unfolding of proteins and nucleic acids. The vast majority of data was downloaded from the *ProTherm* database [3,4] at http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html and controlled in the original literature. For each dataset $T_{ref} = T_{\Delta H = T \cdot \Delta S} = T_m$. T_m is the so-called midpoint or equilibrium temperature, the temperature at which in a dynamic and fully reversible two-state equilibrium the fractions of both (two particularly stable and well observable) macrostates are equal, therefore $\Delta G_{T_m} = 0$ (eqn. 3). We

expanded the above datasets with an additional function each, the state function $\Delta G_T = \Delta H_T - T \cdot \Delta S_T$, using equations 3 (right-hand side), 6 and 7. At that stage, no numerical values were attributed to T yet. Each dataset was now made up of five 'characterizing parameters' $\{\Delta C_p; \Delta H_{T_m}; \Delta S_{T_m}; T_m = \Delta H_{T_m}/\Delta S_{T_m}; \Delta G_T = \Delta H_T - T \cdot \Delta S_T\}$, *all of which are dependent on one another through the fundamental thermodynamic equations 1 to 5*, and of one 'independent variable' T. Note that all five parameters, despite being derived from C_p and T, bear distinct physical meanings (interpretations).

All 1555 datasets were then grouped into 154 families, according to the structural similarity of the members within each group (mostly 'single-chain link' variants, 'point mutants'). The datasets of each of the 154 groups were submitted to a group-specific correlation between the two combined (with respect to ΔH and ΔS) parameters ΔG_T and T_m. An increasingly refined sampling of ΔG_T on a representative part of the groups led to a complete correlation analysis $\Delta G_{T_{median}}$ vs. T_m of all groups at a group-specific $T = T_{median}$. T_{median} is the statistical median of all equilibrium temperatures T_m of a group.

Results

The correlations between T = 273 and 373 K appeared visibly linear for the vast majority of the analysed groups, hence, a linear regression according to equation 8 was used to characterise every group.

$$\Delta G_T = h_T - T_m \cdot s_T = h_T - (\Delta H_{T_m}/\Delta S_{T_m}) \cdot s_T \qquad (8)$$

Detailed results are described in the additional files 1 and 2. Here it suffices to note that all members of the same group share the same 'group parameters' h_T and s_T which express nothing more than the average energy and, respectively, entropy of the group of similar objects. They are therefore only dependent on T and the choice of which individual members constitute 'a group'. The numerical values for the slope $s_{T_{median}}$ are actually average values of all numerical ΔS_{T_m} values of each group member within one group. The numerical values for h_T and all other s_T depend on $\Delta C_p(T)$, the more so the larger $|T - T_{median}|$ is. According to equation 8 the T-dependence of h_T and s_T is the same as for ΔG_T. For ΔC_p = const. this T-dependence adopts the form $f(T) = a + b \cdot T + c \cdot T \cdot \ln T$, in which c is nil for $\Delta C_p = 0$ (eqns. 3, 6 and 7). We fitted this function to all experimental data, to obtain the 'group constants' (with respect to T) h_{0-2} and s_{0-2} for $h_T = h_0 + h_1 \cdot T + h_2 \cdot T \cdot \ln T$ and $s_T = s_0 + s_1 \cdot T + s_2 \cdot T \cdot \ln T$. Note that h_{0-2} and s_{0-2} [see additional file 2] can all be derived from the ΔS_{T_m}, ΔC_p and T_m values of a group with no additional information or assumptions (eqn. 42 [see additional file 1]).

The main result is that at T_{median}, at the temperature where the sum of ΔG of all group members within one group is closest to nil, the vast majority of experimental data produces a linearity of unexpected quality. The linearity as such remains visible but its quality, as expressed through the regression coefficient, degrades quite strongly and monotonously with increased $|T - T_{\mathrm{median}}|$ (Figures S14-S15 [see additional file 1]) and, in a non-trivial fashion, as we join evidently less similar objects into the analysed group (Figures S1, S5-S6, S10-S11 [see additional file 1]). The experimental group sizes vary between 4 and 68 (average 10). The regression coefficients $r_{T_{\mathrm{median}}}$ of all calorimetric groups lie between 0.90 and 0.999'999 with an abundance maximum between 0.999 and 0.9999 (Figures S12-S13 [see additional file 1]). The van't Hoff groups do not fall far behind (Figure S7 [see additional file 1]). In addition, the same correlation method was tested on the calculated thermodynamics of formation from the pure chemical elements in their standard state of a homologue series of PM3-calculated simple organic molecules, as well as of published ab initio-calculated water clusters [5], using statistical thermodynamics at 298 K. The somewhat lower correlation coefficients r_{298K} as compared to the above experimental $r_{T_{\mathrm{median}}}$ values are due to the fact in part that at T = 298 K many calculated datapoints within one group do not center around ΔG = 0. The linearity of similar groups is nevertheless unambiguously apparent (Figures S37-S39 [see additional file 1]).

Discussion

The mere fact that changes in energy and entropy are fundamentally correlated is not unexpected; after all, their temperature dependence is akin and dictated by the corresponding change in heat capacity (eqn. 1), i.e., their mean fluctuation (eqn. 2). A relationship between free energy and the temperature at which it vanishes is not astonishing either. Both ΔG_T and T_{m} are commonly interpreted as a representation of 'thermodynamic stability', the former is expressed in energy units and depends on $\Delta C_p(T)$, the latter lends its unit from the temperature scale and is untouched by any T-dependence of ΔC_p. However, we were unable to find in the literature any systematic study that would demonstrate this particular linearity from experimental data, nor its strong dependence on the similarity of congeners, nor its highest quality at $T = T_{\mathrm{median}}$. The distinct linear grouping of the theoretically calculated molecules (of chemically very different nature from that of proteins or nucleic acids) is at least inasmuch significant as their thermodynamic parameters are independently derived from partition functions rather than from experimental enthalpies or experimental equilibrium constants, and in
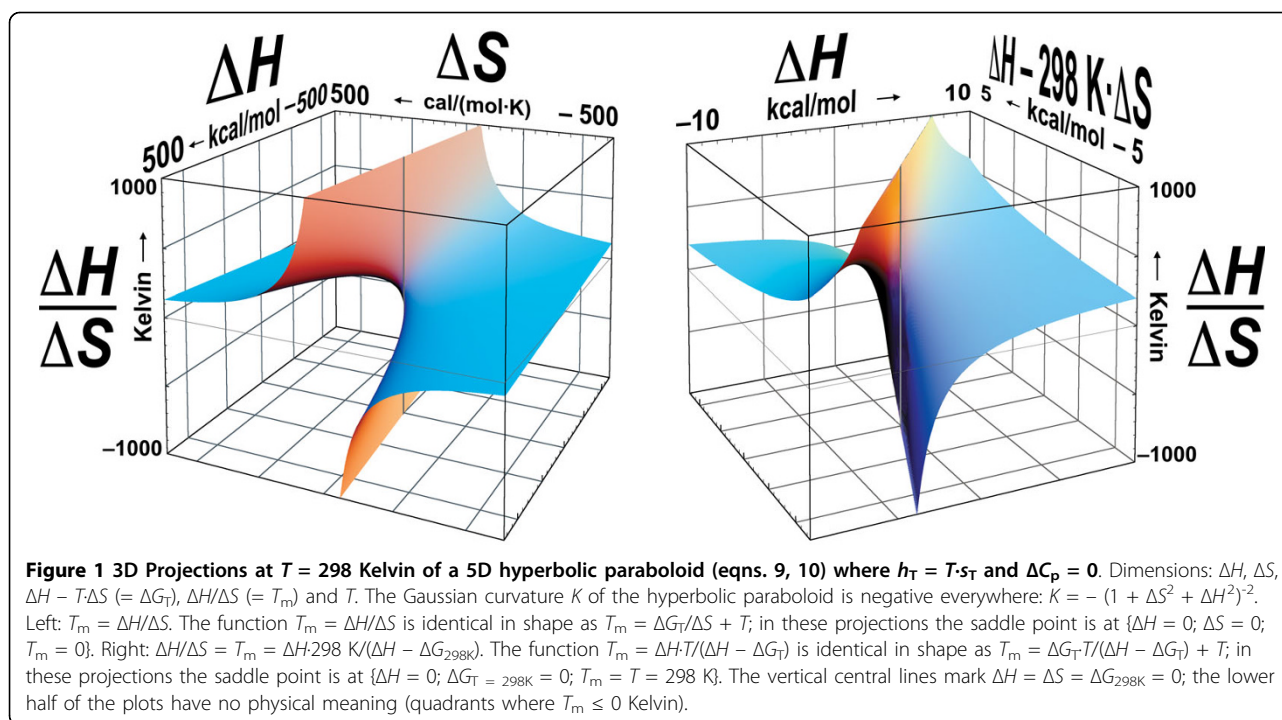
spite of the not entirely exact nature of the calculation of S (due to the harmonic oscillation approximation).

Taken together, the similarity-dependent linearity of $\Delta G_{T_{\mathrm{median}}}$ vs. T_{m}, quantified through the regression coefficient $r_{T_{\mathrm{median}}}$, seems to be as general as the whole theory of thermodynamics is. It may thus be that this linearity's origin lies at least in part in the mathematical structure of thermodynamics, not entirely in the physics for which thermodynamics was designed to describe. Therefore we proceed with deriving general consequences, with respect to physics, such as the entanglement of the First and Second Laws for groups of similar objects as mentioned in the introduction. We continue with the mathematical and geometrical analysis of a function that was generated from the combination of equations 3, 8 (both right-hand side), 4 and 5 to give through the elimination of ΔG_T equations 9 and 10, i.e., the fundamental energy-entropy relationship and mathematical basis for the 5D parameter space $\{\Delta H_{T_{\mathrm{m}}}; \Delta S_{T_{\mathrm{m}}}; T_{\mathrm{m}} = \Delta H_{T_{\mathrm{m}}} / \Delta S_{T_{\mathrm{m}}}; \Delta G_T = \Delta H_T - T \cdot \Delta S_T; T\}$. Equation 9 is a simplified version for $\Delta C_p = 0$ (for clarity) of the general form as shown in equation 10. Both equations can be analytically solved for $\Delta S_{T_{\mathrm{m}}}$ (eqn. 26 [see additional file 1]).

$$\Delta H_{T_{\mathrm{m}}} = T \cdot \Delta S_{T_{\mathrm{m}}} \cdot \frac{\dfrac{h_T}{T} + \Delta S_{T_{\mathrm{m}}}}{s_T + \Delta S_{T_{\mathrm{m}}}} \tag{9}$$

$$\Delta H_{T_{\mathrm{m}}} = T \cdot \Delta S_{T_{\mathrm{m}}} \cdot \frac{\dfrac{h_T \displaystyle\int_{T_{\mathrm{m}}}^{T} \Delta C_p(T)\mathrm{d}T + T \cdot \displaystyle\int_{T_{\mathrm{m}}}^{T} \left(\dfrac{\Delta C_p(T)}{T}\right)\mathrm{d}T}{T} + \Delta S_{T_{\mathrm{m}}}}{s_T + \Delta S_{T_{\mathrm{m}}}} \tag{10}$$

The above functions are variants of the well known quadric $x = y \cdot z$ of the shape of a hyperbolic paraboloid (where $x = \Delta H_{T_{\mathrm{m}}}$, $y = \Delta S_{T_{\mathrm{m}}}$ and $z = T$), thus, of a single saddle point centered in the origin $\{x = 0; y = 0; z = 0\}$ and the S_4-symmetric function spreading from there with an all-negative Gaussian curvature (Figure 1). Any temperature dependence of $\Delta C_p(T)$ is consistent with the hyperbolic paraboloid (eqn. 9) as shown in equation 10. For $\Delta C_p = 0$ (eqn. 9 with $h_T = h_0 + h_1 \cdot T$ and $s_T = s_0 + s_1 \cdot T$ from the van't Hoff datasets) the basic shape of the function does not change when compared to $x = y \cdot z$, although the function area may be quite heavily 'distorted' (not shown). However, for $\Delta C_p \neq 0$ = const. (eqn. 9 with $h_T = h_0 + h_1 \cdot T + h_2 \cdot T \cdot \ln T$ and $s_T = s_0 + s_1 \cdot T + s_2 \cdot T \cdot \ln T$) the group constants h_{0-2} and s_{0-2} that were obtained from the experimental calorimetric datasets produced shapes of the eyebrow-rising kind. In Figure 2 four views of the same 3D-projection, ΔH_T versus ΔS_T and T, of the thermodynamic 5D parameter space

Figure 1 3D Projections at $T = 298$ Kelvin of a 5D hyperbolic paraboloid (eqns. 9, 10) where $h_T = T \cdot s_T$ and $\Delta C_p = 0$. Dimensions: ΔH, ΔS, $\Delta H - T \cdot \Delta S$ (= ΔG_T), $\Delta H / \Delta S$ (= T_m) and T. The Gaussian curvature K of the hyperbolic paraboloid is negative everywhere: $K = -(1 + \Delta S^2 + \Delta H^2)^{-2}$. Left: $T_m = \Delta H / \Delta S$. The function $T_m = \Delta H / \Delta S$ is identical in shape as $T_m = \Delta G_T / \Delta S + T$; in these projections the saddle point is at $\{\Delta H = 0; \Delta S = 0; T_m = 0\}$. Right: $\Delta H / \Delta S = T_m = \Delta H \cdot 298\ K / (\Delta H - \Delta G_{298K})$. The function $T_m = \Delta H \cdot T / (\Delta H - \Delta G_T)$ is identical in shape as $T_m = \Delta G_T T / (\Delta H - \Delta G_T) + T$; in these projections the saddle point is at $\{\Delta H = 0; \Delta G_{T = 298K} = 0; T_m = T = 298\ K\}$. The vertical central lines mark $\Delta H = \Delta S = \Delta G_{298K} = 0$; the lower half of the plots have no physical meaning (quadrants where $T_m \leq 0$ Kelvin).
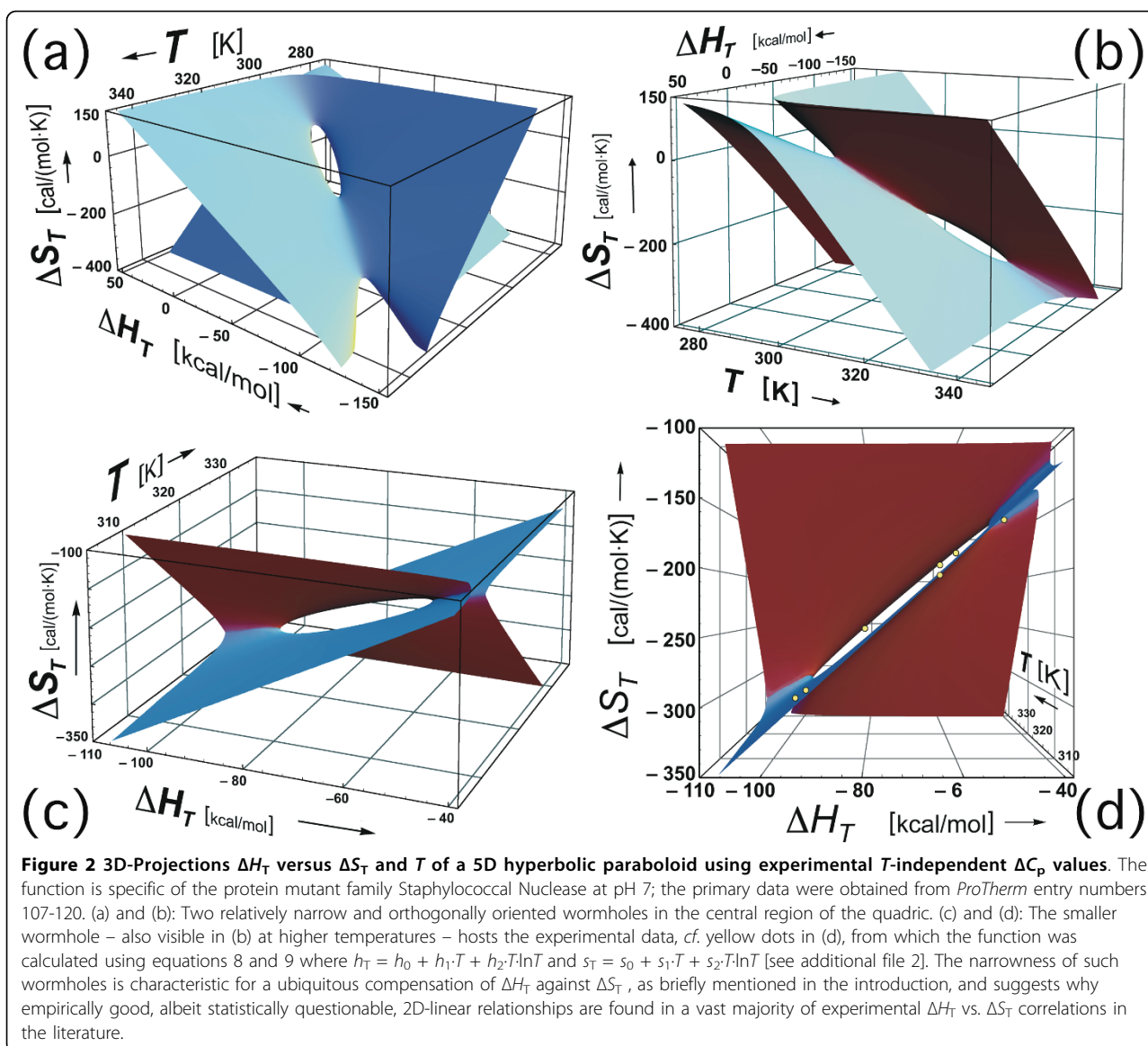
is shown for one particular but representative calorimetrically measured protein mutant group (mutants of Staphylococcal Nuclease). In Figure 3 one to two views of three different 3D-projections for the same mutant group are depicted. Both Figures 2 and 3 focus on the zone that contains the experimental data (yellow dots). The interested reader is welcome to copy any set of experimental group constants h_{0-2} and s_{0-2} [additional file 2], plot equation 9 at any scale (best solved for ΔS_{T_m} to suppress a maximum of asymptotic planes in certain 3D projections) and enjoy the shapes and wormholes created by the $T \cdot \ln T$ terms. A more comprehensive study on the characteristics of this function shall be published elsewhere.

The yellow line in Figure 3d, i.e. the experimental isotherm at $T = T_{median}$, lies in a 'valley' at $T_{median} = 320.2$ Kelvin created by the saddle of this particular hyperbolic paraboloid. It seems that this isotherm is the best defined of all T, therefore, producing the best linear regression coefficient $r_{T_{median}}$. Each straight line in ΔG_T versus $(\Delta H / \Delta S)_{T_{\Delta G=0}}$ that represents a structurally similar group is, in geometric terms, a geodesic on the hyperbolic paraboloid. The corresponding group functions $\Delta H_{T_m}(\Delta S_{T_m})$ or $\Delta S_{T_m}(\Delta H_{T_m})$, as expressed through equations 9 and 10 are therefore also geodesics. Geometric considerations indicate that the datapoints produce the best r_T values in $\Delta G_{T_{median}}$ vs. $(\Delta H / \Delta S)_{T_{\Delta G=0}}$ when they are closest to the maximal negative curvature, thus, to the saddle point of the hyperbolic paraboloid (*cf.* Figure 3d). Flatter curvatures,

thus, steeper surface areas of the hyperbolic paraboloid farther away from the saddle point (*cf.* Figure 1) allow for a higher dispersal of the datapoints owing to idiosyncratic ΔC_p values, which leads to lower regression coefficients r_T.

Independently of geometric considerations, we interpret this consistently observed linearity as a (physically) 'minimal expense' or (mathematically) 'minimal action' effect: The appearance or evolution of small structural changes within the same group, i.e., without touching essential framework structuring, can only result in constantly proportional, therefore, unevolving free energy changes being 'linear' with respect to their equilibrium temperature changes. A thermodynamic interpretation of this linear relationship would be that incremental irreversible changes within a group of reversibly dynamic similar but distincty different structures are just as reversible changes are: virtually uncoupled, therefore, additive and independent of the path taken in between, as is the prerequisite for obeying the Gibbs-Helmholtz equation and synonymous to ΔG and ΔF being state functions.
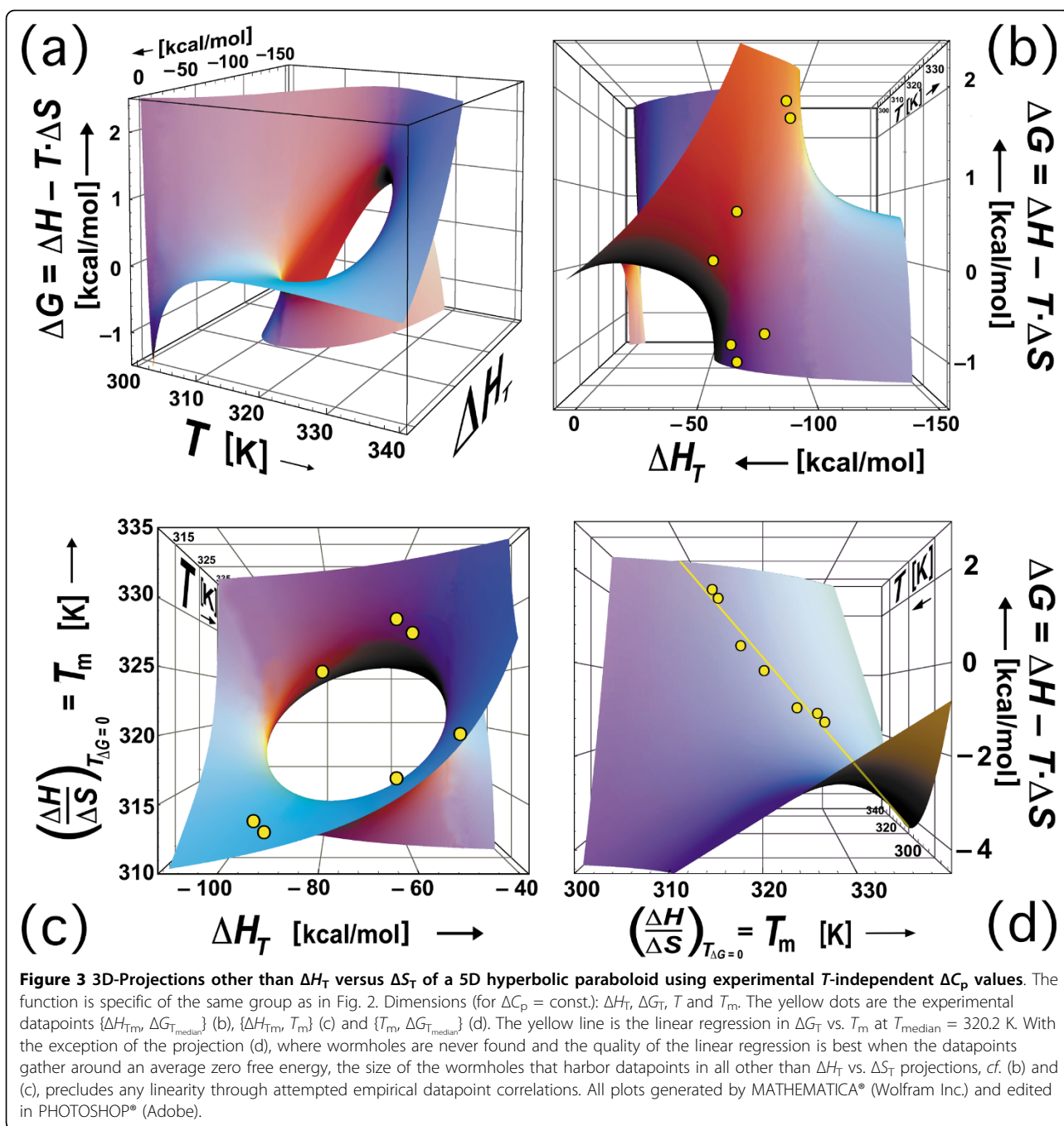
One might argue that the linearity of equation 8 is a simplified manifestation of the Taylor series expansion for any mathematical function $f(x) = f(x_0) + (df/dx) \cdot (x - x_0) + (d^2 f/dx^2) \cdot (x - x_0)^2 + (d^3 f/dx^3) \cdot (x - x_0)^3 + \dots$ which always becomes approximately linear for any slowly varying function $f(x)$, $\Delta G_{T_{median}}$ in this case, sufficiently close to the reference point x_0 (T_m or $\Delta H_{T_m} / \Delta S_{T_m}$ in this case). In performing the linear correlations ΔG_T

Figure 2 3D-Projections ΔH_T versus ΔS_T and T of a 5D hyperbolic paraboloid using experimental T-independent ΔC_p values. The function is specific of the protein mutant family Staphylococcal Nuclease at pH 7; the primary data were obtained from *ProTherm* entry numbers 107-120. (a) and (b): Two relatively narrow and orthogonally oriented wormholes in the central region of the quadric. (c) and (d): The smaller wormhole – also visible in (b) at higher temperatures – hosts the experimental data, *cf.* yellow dots in (d), from which the function was calculated using equations 8 and 9 where $h_T = h_0 + h_1 \cdot T + h_2 \cdot T \cdot \ln T$ and $s_T = s_0 + s_1 \cdot T + s_2 \cdot T \cdot \ln T$ [see additional file 2]. The narrowness of such wormholes is characteristic for a ubiquitous compensation of ΔH_T against ΔS_T, as briefly mentioned in the introduction, and suggests why empirically good, albeit statistically questionable, 2D-linear relationships are found in a vast majority of experimental ΔH_T vs. ΔS_T correlations in the literature.

versus $\Delta H_{T_m} / \Delta S_{T_m}$ at $T = T_{median}$, we do not explicitly claim that the linear relation holds at all temperatures. We do claim, however, that a correlation between ΔG_T and T_m at any temperature T using a polynomial of higher than first (linear) degree, as generalised in the above Taylor series expansion, will lead to an analytically solvable relationship for $\Delta H_{T_m}(\Delta S_{T_m})$ or $\Delta S_{T_m}(\Delta H_{T_m})$. We did not prove the generality of this claim but solved $\Delta H - T \cdot \Delta S = h_T - [(\Delta H/\Delta S) \cdot s_{1,T} + (\Delta H/\Delta S)^2 \cdot s_{2,T} + (\Delta H/\Delta S)^3 \cdot s_{3,T}]$, which is a Taylor series-expanded version of equation 8 (where $\Delta C_p = 0$), for ΔH and ΔS, respectively. The expanded nonlinear variants with $s_{3,T} = 0$ (quadratic) and $s_{3,T} \neq 0$ (cubic) did each result in at least one non-complex analytical solution for $\Delta H(\Delta S)$ and $\Delta S(\Delta H)$, albeit bearing a more complicated mathematical structure (not shown).

In other words, we claim that *a fundamental relationship between energy and entropy for a group of similar objects results from **any** analytically solvable relationship between ΔG_T and $\Delta H_{T_m}/\Delta S_{T_m}$*. We opt for the simplest, a linear solution: ΔG_T and $\Delta H_{T_m}/\Delta S_{T_m}$ are proportional over a reasonably large temperature range.

Most important for physics is the fact that group specific thermodynamic parameter spaces depict the only possible values that can be realised by a particular group of similar objects. The rest is void, *terra incognita* for the group members, unless an object changes its characteristics (structure, composition, etc.), unless it 'dissimilarises' off from 'its' group - most likely, to join some other one. The definition of a group, that is, how to determine whether a number of individuals belong to the same group or not, seems at first sight worrying or

Figure 3 3D-Projections other than ΔH_T versus ΔS_T of a 5D hyperbolic paraboloid using experimental *T*-independent ΔC_p values. The function is specific of the same group as in Fig. 2. Dimensions (for ΔC_p = const.): ΔH_T, ΔG_T, T and T_m. The yellow dots are the experimental datapoints $\{\Delta H_{Tm}, \Delta G_{T_{median}}\}$ (b), $\{\Delta H_{Tm}, T_m\}$ (c) and $\{T_m, \Delta G_{T_{median}}\}$ (d). The yellow line is the linear regression in ΔG_T vs. T_m at T_{median} = 320.2 K. With the exception of the projection (d), where wormholes are never found and the quality of the linear regression is best when the datapoints gather around an average zero free energy, the size of the wormholes that harbor datapoints in all other than ΔH_T vs. ΔS_T projections, *cf.* (b) and (c), precludes any linearity through attempted empirical datapoint correlations. All plots generated by MATHEMATICA® (Wolfram Inc.) and edited in PHOTOSHOP® (Adobe).

at least not clearly solved. However, when we think of individuals as being more or less similar to one another, we see that a clear distinction between different groups is not a fundamental issue. Similarity does exist; in the microscopic and macroscopic world it is often a matter of judgement according to some objective, statistically relevant technical signal (at highest available resolution) or at least a subjective physiological 'measurement' ("I know it when I see it", *cf.* Graphical Abstract). For microscopic objects such as molecules, one should

never be tempted to define a group through a good linear regression coefficient only; independent knowledge and/or studies are mandatory. For instance, the advantage of studying mutant protein families not only means being able to analyse a large number of families and sometimes many congeners within one family. Most importantly, we are also certain that single or even multiple site mutants of the same protein do indeed belong to the same structural group, the mutants are undoubtedly similar to one another. Other molecular systems

such as synthetic host-guest complexes or water clusters may be less evident to this respect. Still other objects might be even more readily grouped than mutant proteins (*cf.* Conclusion). The concept of similarity is intrinsically a not readily quantifyable one because intuitively it seems to be a not very objective 'measurement', at least down to Planckian scales: How similar and with respect to what exactly?

We are free to group similar objects essentially at will. For example, we can group one set of RNA hairpins into two families, the one that bears various all-Watson-Crick pairs and the one that contains various single-mismatched base pairs at different positions in the stem, the stem length and loop sequence being the same in both families [6]. We can overlook this subtle difference and treat those hairpins as one group that consist of the same loop sequence and stem length irrespective of single mismatches being present or absent in the stem. The outcome will be a slightly lower linear regression coefficient for this group. It can then be compared to another group of RNA hairpins showing, for example, the same stem length and stem sequence variations but a different loop sequence. We can treat protein mutant families with the same varied degrees of precision/resolution. We could define all known proteins as belonging to the same group and compare it to a more drastically different group of compounds (objects). Nothing prevents us from grouping objects at still lower resolution; the obvious trade-off will be increasingly lower linear regression coefficients. As a matter of fact, there is no *a priori* objection that we can think of to the grouping of the entire universe and comparing it to some other one, if it were observable. In principle, one would have to agree upon a set of observables (like energy, entropy and temperature), measure them on a statistically representative number of individual members of what we decide, through some hopefully objective criterium, to call a group, determine the corresponding group parameters and then gain easier access to more members of the same group but also, to obtain an objective means for the comparison of this group to another one. In practise, of course, as we embrace more and more dissimilar objects, we will probably evoke increasingly unacceptable linear regression coefficients. Where this limit of a meaningful group analysis lies remains to be seen.

Conclusion

In this study we introduce a geometrical parameter space description of thermodynamics and offer a general way of objectively quantifying similarity (to whatever resolution) of individual objects based on two well known abstract notions (not postulated 'empirical' physical parameters): the use of the knowledge of a group membership, and the mathematical relationship between difference and ratio being the results from the two most fundamental mathematical operations, substraction and, respectively, division. The latter notion opens access to a higher than three-dimensional (ΔH, ΔS, T) geometrical description of thermodynamics through expansion of the parameter space with $\Delta H - T \cdot \Delta S$ and $\Delta H / \Delta S$. The combination of both notions indicates a group-related redundancy in the mathematical structure of thermodynamics; a redundancy which becomes evident when relating substraction and division for the characterisation of similar objects. This redundancy necessarily unravels a group-related fundamental relationship between energy and entropy for similar objects and, possibly, a general unified law of thermodynamics for structured matter. According to our findings, any group of similar objects may be characterised by precisely how the energy and entropy of each individual group member is related (coupled) to one another. We show that similar dynamic structures, for example molecules, 'minimise their action' on thermodynamic state changes such that, within a structural framework — within 'a group' as specified by the group parameters h_T and s_T using equations 8, 9 and 10 — the distinction between energy and entropy becomes a formal one.

The usually incomplete knowledge of all molecular properties of a thermodynamic system, such as differential solvation, salt, and bulk solvent effects in biomolecular systems, continues to confront us with the limitation of exactly calculating the free energy, the enthalpy, or the entropy from the fundamentals. However, having at hand reliable experimental or theoretical data of both ΔG and ΔH of as many group members of similar structures as possible, thus, of a statistically sufficient number of group members, we can predict from either ΔH or ΔG of more group members their respective ΔG or ΔH and concurrently ΔS. The relatively simple mathematical structure of group thermodynamics allows us to quantify through linear regressions the structural similarity imprinted into the thermodynamic behavior of, in principle, any structural framework. On a molecular scale, group thermodynamics may strongly simplify the elucidation of entropies of molecules that are known to belong to a group of similar compounds through a bypass of costly calculations of the vibrational components of idealised partition functions. With the knowledge of the group parameters h_T and s_T at hand, S can be calculated from U or H. In addition, it may be a possibly useful complement for cross-checking ΔG calculations that have been obtained from simulations using molecular dynamics techniques. Generally group thermodynamics may contribute to systematic analyses in biomolecular and chemical thermodynamics and, when applied to chemical reaction kinetics, in systems chemistry.

Theories from quite different domains such as, to name a few, probability theory [7-10], information theory and the emergence of complex systems [11-18], quantum relativity/cosmology [19-29] and string theory [30] operate with entropy and the Second Law of thermodynamics yet in conjunction with parameters different from the ones studied here. Urgent problems are being at least attacked, and possibly solved, through the insight into apparent and/or fundamental analogies between statistical thermodynamics and, for example (respectively), randomness of sequential irregularities ("algorithmic entropy", "approximate entropy"), computational compactness ("logical depth"), quality change of hereditary information (change in systemic "knowledge" through periodically discarded "Shannon entropy"), the dynamics of black holes ("Bekenstein-Hawking entropy"), and tracing back the microscopic origin of their area-entropy by counting the degeneracy of periodic and persistent topological defects (Bogomol'nyi-Prasad-Sommerfield soliton bound states) in certain kinds of supersymmetric branes that mimic the thermodynamics of idealised extremal, highly charged black holes. In all above cases the problem arises of how to reliably quantify or sample randomness, logical depth, knowledge, entropy, in order to understand their physical origins and perhaps their development over time. The energy-entropy relationship derived from thermodynamic group characteristics may help solve one or the other problem, in particular, when the to be analysed physical objects are not as potentially overwhelmingly dissimilar as chemical systems can be — in order to ease, for a start, the choice of groups.

Black holes, being the most immensely dense and, with respect to their composition, the perhaps most uniform objects known in physics, are all in a state of maximal entropy and are thought to differ from one another through, out of all known matter, the least of characterising parameters; only mass, angular momentum and, for some limited time period, electric charge makes them different: "black holes have no hair". In contrast, elementary particles may differ through a whole plethora of characteristics (according to the standard model) and the variability, thus, potential dissimilarity of objects that are composed of these elementary particles (of 'normal' nonrelativistic matter) multiplies, i.e., increases at a geometric rate with the number of involved particles. If micro black holes indeed existed and could be transiently generated in future Large Hadron Collider experiments, if different classes of such potentially highly similar objects could be observed and analysed, we would predict that the relationship between their gravitational energy and the surface area of their event horizon would correlate in a fashion that were characteristic for their kind: Energy (= mass) and entropy (= surface)

would correlate, through equation 10, differently, i.e., with different group parameters for objects of a particular (range of) angular momentum and electric charge than for another. Distinct groups should appear and be best visible in free energy correlations as formulated in equation 8. A difficulty might arise from the fact that micro black holes are not expected to be formed in a thermodynamic equilibrium, but rather 'kinetically controlled'. How then to measure free energy? We imagine that a measure of free energy of micro black holes would be their abundance under given experimental conditions: Plot under maximum and constant total abundance ('steady state') conditions the logarithm of abundance (through counting) versus ratio of gravitational energy (mass) over surface (of the event horizon). The linearity should produce the best linear regression coefficients when, within a group of analysed micro black holes, the median mass is populated most.

Additional material

Additional file 1: GraphMath_SI. Graphs containing a large number of representative regression plots, statistical analyses and the Mathematical Appendix.

Additional file 2: NumSI. Numerical primary data (tab-delimited), optimised parameters and regression coefficients from linear regressions and non-linear curve fittings, which can be independently readily reproduced from the given primary data.

Authors' contributions

PZ derived the Mathematical Appendix [see additional file 1] and contributed significantly to the correct description of the mathematical relationships (in particular, eqn. 10 and *T*-dependence of C_p, h_T and s_T) and much of the fundamental physics in the text. ST extracted all primary data from the *ProTherm* and *ProNIT* databases at http://gibk26.bse.kyutech.ac.jp/jouhou/, cross-checked the numerical values and analysed all error margins in the original literature, carried out [see additional file 2] and plotted all linear regressions and polynomial fittings (Figures S2, S3, S4, S8, S9 [see additional file 1]). PS derived equations 8, 9 and $\Delta S_{T_m}(\Delta H_{T_m})$ as shown in

equation 26 [see additional file 1], conceived of the study and wrote the manuscript and both additional files. All authors read and approved the final manuscript and both additional files.

Competing interests

The authors declare that they have no competing interests.

References

1. Prabhu NV, Sharp K: **Heat capacity in proteins.** *Annu Rev Phys Chem* 2005, **56**:521-48.
2. Benzinger TH: **Thermodynamics, chemical reactions and molecular biology.** *Nature* 1971, **229**:100-2.
3. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A: **ProTherm, version 4.0: thermodynamic database for proteins and mutants.** *Nucleic Acids Res* 2004, **32**:D120-21.
4. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A: **ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions.** *Nucleic Acids Res* 2006, **34**:D204-6.
5. Dunn ME, Pokon EK, Shields GC: **Thermodynamics of Forming Water Clusters at Various Temperatures and Pressures by Gaussian-2, Gaussian-3, Complete Basis Set-QB3, and Complete Basis Set-APNO Model Chemistries; Implications for Atmospheric Chemistry.** *J Am Chem Soc* 2004, **26**:2647-53.
6. Strazewski P: **Thermodynamic Correlation Analysis: Hydration and Perturbation Sensitivity of RNA Secondary Structures.** *J Am Chem Soc* 2002, **124**:3546-54.
7. Chaitin GJ: **Randomness in arithmetic.** *Sci Am* 1988, **259**:80-5.
8. Pincus SM: **Approximate entropy as a measure of system complexity.** *Proc Natl Acad Sci USA* 1991, **88**:2297-301.
9. Pincus S, Singer BH: **Randomness and degrees of irregularity.** *Proc Natl Acad Sci USA* 1996, **93**:2083-88.
10. Pincus SM, Kalman RE: **Irregularity, volatility, risk, and financial market time series.** *Proc Natl Acad Sci USA* 1997, **101**:13709-14.
11. Kuhn H: **Model Consideration for the Origin of Life.** *Naturwissenschaften* 1976, **63**:68-80.
12. Bennett CH: **On the nature and origin of complexity in discrete, homogeneous, locally-interacting systems.** *Found Phys* 1986, **16**:585-92.
13. Bennett CH: **Information, Dissipation, and the Definition of Organization.** In *Emerging Syntheses in Science.* Edited by: Pines D. Addison-Wesley, Massachusetts; 1987:297.
14. Kuhn H: **Origin of life and physics: Diversified microstructure - Inducement to form information-carrying and knowledge-accumulating systems.** *IBM J Res Devel* 1988, **32**:37-46.
15. Lloyd S, Pagels H: **Complexity as Thermodynamic Depth.** *Ann Phys* 1988, **188**:186-213.
16. Landauer R: **A simple measure of complexity.** *Nature* 1988, **336**:306-7.
17. Kuhn H: **Origin of life - Symmetry breaking in the universe: Emergence of homochirality.** *Curr Op Colloid Interface Sci* 2008, **13**:3-11.
18. Kuhn H: **Is the transition from chemistry to biology a mystery?** *J Syst Chem* 2010, **1**:3.
19. Christodolou D: **Reversible and irreversible transformations in black-hole physics.** *Phys Rev Lett* 1970, **25**:1596-97.
20. Christodolou D, Ruffini R: **Reversible transformations of a charged black hole.** *Phys Rev* 1971, **D4**:3552-55.
21. Penrose R, Floyd R: **Extraction of rotational energy from a black hole.** *Nature Phys Sci* 1971, **229**:177-9.
22. Hawking SW: **Gravitational radiation from colliding black holes.** *Phys Rev Lett* 1971, **26**:1344-6.
23. Bekenstein JD: **Black holes and the second law.** *Nuovo Cimento Lett* 1972, **4**:737-40.
24. Bekenstein JD: **Black holes and entropy.** *Phys Rev* 1973, **D7**:2333-46.
25. Bekenstein JD: **Generalized second law of thermodynamics in black-hole physics.** *Phys Rev* 1974, **D9**:3292-300.
26. Carter B: **Rigidity of a black hole.** *Nature* 1972, **238**:71-2.
27. Bardeen J, Carter B, Hawking S: **The four laws of black hole mechanics.** *Comm Math Phys* 1973, **31**:161-70.
28. Hawking SW: **Black hole explosions?** *Nature* 1974, **248**:30-1.
29. Hawking SW: **Particle creation by black holes.** *Comm Math Phys* 1975, **43**:199-220.
30. Strominger A, Vafa C: **Microscopic origin of the Bekenstein-Hawking entropy.** *Phys Lett B* 1996, **379**:99-104 [http://arxiv.org/abs/hep-th/9601029v2].