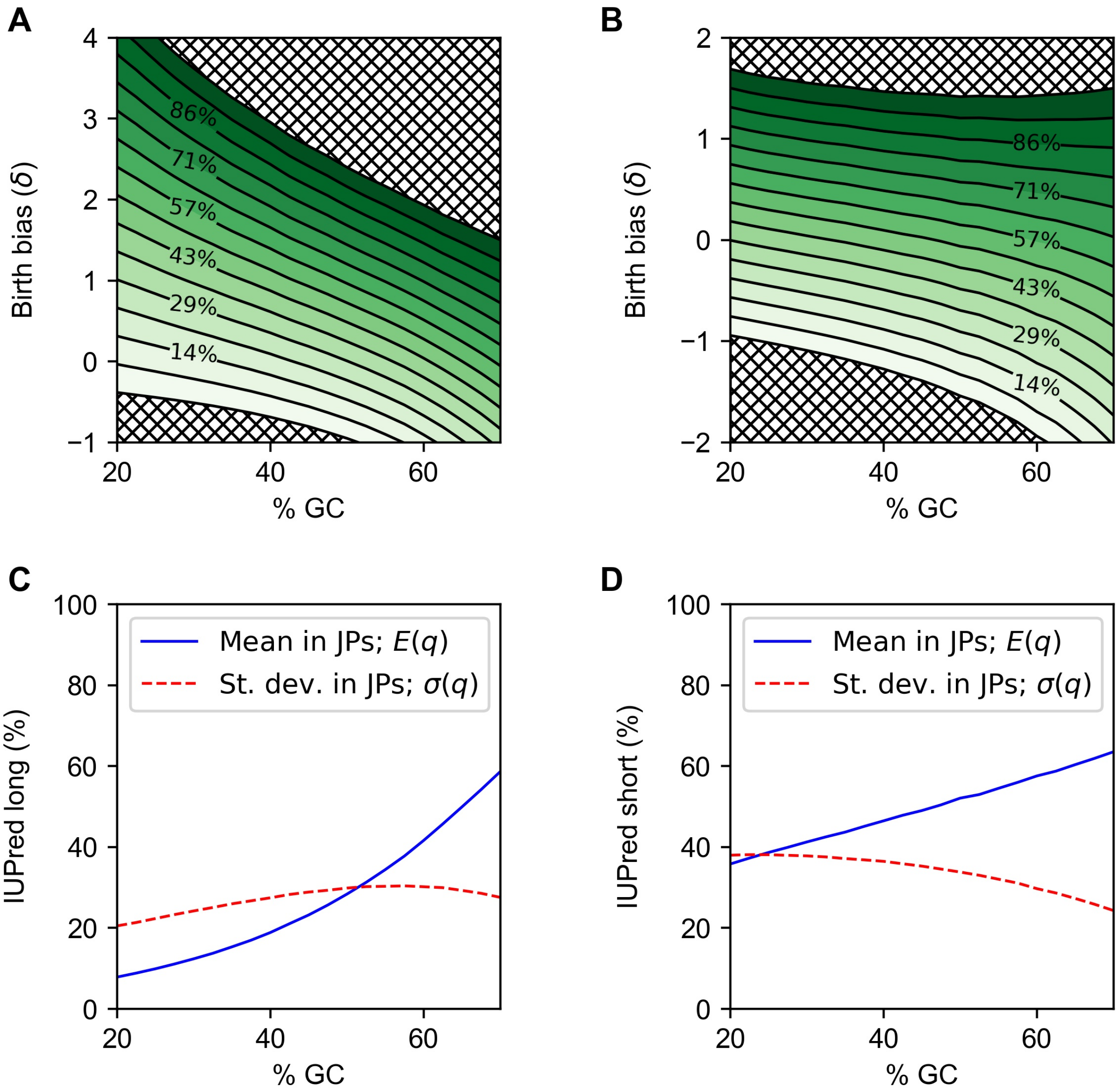

Figure S1. IUPred "long" and "short" estimators of intrinsic structural disorder disagree on the relation between GC content and the intrinsic structural disorder of junk polypeptides and novel functional polypeptides under a random-sequence model. (*A*) Contour plot of the predicted average of IUPred "long" disorder among novFPs (identical to fig. 2B). (*B*) Contour plot of the predicted average IUPred "short" disorder among novFPs. (*C*) The predicted mean and standard deviation of IUPred "long" disorder among JPs as functions of the GC content (identical to fig. 2D). (*D*) The predicted mean and standard deviation of IUPred "short" disorder among JPs as functions of the GC content. Hatched areas

indicate impossible percentages of ISD, i.e. outside the 0%-100% interval. The landscapes in panels A and B can be understood as the results of applying equation 2 to the curves in panels C and D, respectively. As a result, the vertical “slice” of a landscape at a given GC content is a straight line whose intercept and slope are respectively the mean and standard deviation associated with this GC content in the corresponding bottom panel. The curve obtained by taking a horizontal “slice” where there is no birth bias ($\delta = 0$) corresponds to the relation between the mean of the property among JPs, i.e. the solid blue curve in the corresponding bottom panel. Since the vertical distance between contour lines is inversely proportional to the vertical slope of the landscape, it is inversely proportional to the standard deviation of the property among JPs, i.e. the dashed red curve in the corresponding bottom panel.

SUPPLEMENTARY METHODS

Applying the Radon-Nikodým theorem to de novo gene birth

This section explains why our framework fits the general setting of the branch of mathematics called measure theory and its sub-branch, probability theory. We introduce some concepts from these theories to clarify why the Radon-Nikodým theorem can be used to compare JPs and novFPs.

Given a set Ω, measure theory provides the basic notions required to develop a self-consistent concept of the “measure” or “size” of subsets of Ω (such as length, area, volume or probability). It is not always possible to consistently define a measure for all subsets of Ω, so that we must choose certain subsets that form a structure called a σ-field (or σ-algebra). A σ-field on Ω is a set $\mathcal{F}$ whose elements are subsets of Ω that meet certain conditions. The consequences of these conditions are that both Ω and the empty set $\emptyset$ are elements of $\mathcal{F}$, and the combination of arbitrary elements of $\mathcal{F}$ through a finite or infinite sequence of standard set operations (union, intersection, complementation, difference and symmetric difference) always produces an element of $\mathcal{F}$ (Vestrup 2003a).

In our framework, the elements of Ω are all the possible polypeptides that are distinct in terms of sequence and/or *cis*-regulation, and the elements of $\mathcal{F}$ are classes of polypeptides. Since the sequence and *cis*-regulatory properties of a polypeptide are determined by a finite DNA sequence containing its ORF, the set Ω is discrete or "countable", i.e. it is not larger than the set of all whole numbers (Komjáth and Totik 2006). Because of this, we can choose $\mathcal{F}$ to be the set of all subsets of Ω, which would cause complications if Ω was a continuum (Vestrup 2003b). Nevertheless, we will continue the explanations in the context of an arbitrary σ-field because that is how the Radon-Nikodým theorem is formulated.

A measure μ defined on a σ-field $\mathcal{F}$ of subsets of Ω is a function that assigns a number to each element of $\mathcal{F}$. If S is an element of $\mathcal{F}$, then $\mu(S)$ denotes the number that μ assigns to S. To meet the definition of a measure, μ must also satisfy three other conditions: 1) $\mu(S) \geq 0$ for each $S \in \mathcal{F}$, 2) $\mu(\emptyset) = 0$, where $\emptyset$ is the empty set, and 3) for any finite or infinite sequence $S_1, S_2, S_3, \ldots$ of non-overlapping elements of $\mathcal{F}$, their union $S = S_1 \cup S_2 \cup S_3 \ldots$ satisfies $\mu(S) = \mu(S_1) + \mu(S_2) + \mu(S_3) + \cdots$ (Vestrup 2003c). The triple $(\Omega, \mathcal{F}, \mu)$ is called a measure space. If a measure P defined on $\mathcal{F}$ also satisfies $P(\Omega) = 1$, then P is called a probability measure and $(\Omega, \mathcal{F}, P)$ is called a probability space, and they are studied by probability theory.

In our framework, we define two probability measures: P, which represents a time average of JPs, and

P_F, which represents novFPs that functionalize in the time period considered. These measures are defined on the same σ-field $\mathcal{F}$; they assign numbers to the same classes of polypeptides. Given S, a subset of Ω which is an element of $\mathcal{F}$, the number $P(S)$ is the ratio of the time-averaged number of JPs that belong to S to the time-averaged total number of JPs. We can see that P meets the three requirements that define a measure: the ratio is never negative ($P(S) \geq 0$), the empty set contains no JPs ($P(\emptyset) = 0$) and the ratio assigned to the union of several non-overlapping classes of polypeptides is the sum of their individual ratios (the numerators add up and the denominator is a constant). P is a probability measure since $P(\Omega) = 1$, i.e. the time-averaged number of JPs that belong to Ω is precisely the time-averaged total number of JPs. P_F is also a probability measure: we define $P_F(S)$ as the proportion of novFPs that belong to S. Proportions are never negative ($P_F(S) \geq 0$), the empty set contains no novFPs ($P_F(\emptyset) = 0$), the proportion of novFPs belonging to the union of several non-overlapping classes is the sum of the proportions belonging to each class, and the proportion of novFPs belonging to the whole set Ω is $P_F(\Omega) = 1$.

Measure theory defines the notion of the integral, with respect to a measure and over a specific subset of Ω, of a numerical function. We use such functions to represent polypeptide properties such as length and intrinsic disorder, and their integrals determine their averages among polypeptides. A function f defined on the set Ω is a function that assigns a number $f(\omega)$ to each element ω of Ω. Given a measure space $(\Omega, \mathcal{F}, \mu)$, a function f on Ω must have a property called $\mathcal{F}/\mathcal{B}^*$-measurability in order for its integral to be well-defined. f is said to be $\mathcal{F}/\mathcal{B}^*$-measurable if, for every real number x, there is an element of $\mathcal{F}$ (called $f^{-1}((x, +\infty]))$ that is exactly the set of all elements ω of Ω which satisfy $f(\omega) > x$ (Vestrup 2003d). Given an $\mathcal{F}/\mathcal{B}^*$-measurable function $f$and a subset S of Ω which is an element of $\mathcal{F}$, the integral of f over S with respect to μ is a number denoted by $\int_S f \, d\mu$. Given a probability space $(\Omega, \mathcal{F}, P)$, the conditional average of f "knowing" S is given (Çinlar 2011) by:

$$E(f|S) = \frac{1}{P(S)} \int_S f \, dP$$

In particular, the (unconditional) average of f is given by:

$$E(f) = E(f|\Omega) = \frac{1}{P(\Omega)} \int_\Omega f \, dP = \int_\Omega f \, dP$$

Given two measures μ and υ defined on the same σ-field $\mathcal{F}$ of subsets of Ω, υ is said to be absolutely continuous with respect to μ if every element S of $\mathcal{F}$ which satisfies $\mu(S) = 0$ also satisfies $\upsilon(S) = 0$. This relationship between μ and υ is also denoted by $\upsilon \ll \mu$ (Vestrup 2003e). In our framework, the measure P represents a time average of JPs and P_F represents novFPs that functionalize in the time period considered. This implies that for each novFP represented in the measure P_F, the JP that it was immediately before functionalization is represented in the measure P. These two polypeptides are identical because of our definition of novFPs, so they belong to exactly the same subsets of Ω. Therefore, if a subset S of Ω is an element of $\mathcal{F}$ and never contains any JPs ($P(S) = 0$), then no novFPs emerge in this subset ($P_F(S) = 0$). Thus, we have $P_F \ll P$.

The Radon-Nikodým theorem for finite measures states that given two measures μ and υ on $\mathcal{F}$ which are both finite ($\mu(\Omega)$ and $\upsilon(\Omega)$ are finite numbers) and which satisfy $\upsilon \ll \mu$, there exists a finite-valued nonnegative $\mathcal{F}/\mathcal{B}^*$-measurable function f on Ω which summarizes the relationship between μ and υ. Specifically, υ can be constructed by integrating f with respect to μ; for each element S of $\mathcal{F}$, we have $\upsilon(S) = \int_S f \, d\mu$ (Vestrup 2003e). In our framework, this theorem applies to the measures P and P_F since they are both finite ($P(\Omega) = P_F(\Omega) = 1$) and $P_F \ll P$. Therefore, there exists a finite-valued nonnegative $\mathcal{F}/\mathcal{B}^*$-measurable function $\hat{r}$ on Ω (a polypeptide property) such that for each element S of

$\mathcal{F}$, we have $P_F(S) = \int_S \hat{r}\, dP$. Because of the definition of the conditional average (Çinlar 2011), we have $P_F(S) = P(S) \times E(\hat{r}|S)$ and thus:

$$E(\hat{r}|S) = \frac{P_F(S)}{P(S)}$$

where $E(\hat{r}|S)$ is the average of $\hat{r}$ among JPs that belong to the class S. This provides an interpretation of the polypeptide property $\hat{r}$: its average among JPs that belong to a given class of polypeptides (S) is the ratio of the frequency of this class among novFPs ($P_F(S)$) to its frequency among JPs ($P(S)$). Since a class of polypeptides may be arbitrarily small and may even contain only one JP, the value of $\hat{r}$ for a single polypeptide is the factor by which its frequency changes from JPs to novFPs. We can deduce from the above equation that the average of $\hat{r}$ among JPs is $E(\hat{r}) = 1$, since:

$$E(\hat{r}) = E(\hat{r}|\Omega) = \frac{P_F(\Omega)}{P(\Omega)} = \frac{1}{1} = 1$$

The function f defined from two measures $\upsilon \ll \mu$ by the Radon-Nikodým theorem has a useful property: for every $\mathcal{F}/\mathcal{B}^*$-measurable function g, its integral with respect to υ over any element S of $\mathcal{F}$ is given by $\int_S g\, d\upsilon = \int_S fg\, d\mu$ (Vestrup 2003e). In our framework, this property translates to $\int_S q\, dP_F = \int_S q\hat{r}\, dP$ for any polypeptide property q. By the definition of the conditional average (Çinlar 2011), we thus have:

$$P_F(S) \times E_F(q|S) = P(S) \times E(q\hat{r}|S)$$

$$E_F(q|S) = \frac{P(S)}{P_F(S)} \times E(q\hat{r}|S)$$

$$E_F(q|S) = \frac{E(q\hat{r}|S)}{E(\hat{r}|S)}$$

where $E_F(q|S)$ is the average of q among novFPs that belong to S. If we choose $S = \Omega$ and use the fact that $E(\hat{r}) = 1$, we obtain:

$$E_F(q) = E(q\hat{r})$$

which shows how the relationship between q and $\hat{r}$ among JPs determines the average of q among novFPs. From this equation, our main mathematical results can be derived using the universal properties of averages, variances, covariances, etc. without further need for the basic concepts of measure theory.

Interpreting the coskewness of three variables

To facilitate the interpretation of the coskewness of three variables $cosk(x, y, z) = \frac{E(\Delta x \Delta y \Delta z)}{\sigma(x)\sigma(y)\sigma(z)}$, where $\Delta x = x - E(x)$, consider the standard score $Z(x) = \frac{\Delta x}{\sigma(x)}$ which has a mean of 0 and a variance of 1.

$$cosk(q, \lambda, f) \ = \ E\big(Z(q)Z(\lambda)Z(f)\big)$$

Since $E(x\ y) = E(x) \times E(y) + cov(x, y)$, we obtain:

$$cosk(q, \lambda, f) \ = E\big(Z(q)\big) \times E\big(Z(\lambda)Z(f)\big) + cov\big(Z(q), Z(\lambda)Z(f)\big)$$

Because $E\big(Z(q)\big) = 0$, we obtain:

$$cosk(q, \lambda, f) \ = cov\big(Z(q), Z(\lambda)Z(f)\big)$$

Because of the definition of coskewness, its value does not change when we swap any two of the three variables:

$$cosk(q, \lambda, f) \ = cov\big(Z(q), Z(\lambda)Z(f)\big) = cov\big(Z(\lambda), Z(q)Z(f)\big) = cov\big(Z(f), Z(\lambda)Z(q)\big)$$

Now consider the fact that $E(Z(x)Z(y)) = cov(Z(x), Z(y)) = \rho(x, y)$. Put in words, the Pearson correlation coefficient is the mean of the product of the standard scores of two variables, while coskewness is the covariance of this same product with the standard score of a third variable.

Therefore, roughly speaking, coskewness is a measure of how any of the three variables linearly affects the correlation between the two others.

LITERATURE CITED

Çinlar, E., 2011 Conditioning, pp. 139-170 in *Probability and Stochastics*, edited by S. Axler and K. A. Ribet. Springer, New York.

Komjáth, P., and V. Totik, 2006 Countability, pp. 9-12 in *Problems and Theorems in Classical Set Theory*, edited by P. Winkler. Springer, New York.

Vestrup, E. M., 2003a Set Systems, pp. 1-34 in *The theory of measures and integration*. John Wiley & Sons, Inc., Hoboken.

Vestrup, E. M., 2003b Lebesgue Measure, pp. 113-162 in *The theory of measures and integration*. John Wiley & Sons, Inc., Hoboken.

Vestrup, E. M., 2003c Measures, pp. 35-61 in *The theory of measures and integration*. John Wiley & Sons, Inc., Hoboken.

Vestrup, E. M., 2003d Measurable Functions, pp. 163-207 in *The theory of measures and integration*. John Wiley & Sons, Inc., Hoboken.

Vestrup, E. M., 2003e The Radon-Nikodym Theorem, pp. 367-436 in *The theory of measures and integration*. John Wiley & Sons, Inc., Hoboken.