Whites Science
A Scientific Erudition

ESTIMATING MAXIMUM PLAUSIBLE CONSERVED SYNTENY BETWEEN ORTHOLOGOUS GENOMES OF A SPECIES PAIR COMPARED IN AN EXHAUSTIVE SEARCH-SPACE

PERAMBUR NEELAKANTA*, SANDHYA SHARMA
Department of Computer and Electrical Engineering & Computer Science Florida Atlantic University, Boca Raton, FL 33431, USA

ABSTRACT

This paper is objectively conceived to find the maximum plausible extent of shared synteny in the chromosomes that could possibly be encountered across an exhaustively-searched, large number of orthologous genomes pertinent to a given pair of (test) species. A viable solution is proposed thereof is as follows: First, by searching a sample-space with a limited number of underlying orthologs of the test species, relevant sparse synteny data (for example, as seen in an Oxford-grid), is gathered and the associated entropy details are ascertained. Then, by judiciously extending such entropy details (however, with the constraint posed by the fixed number of chromosomes in each test species), the maximum synteny that could be encountered when an ensemble of a large number of ortholog pairs is exhaustively searched, is elucidated in terms of appropriately defined metrics. These metrics are derived to specify the maximum extent of plausible shared-synteny, as a function of the number of ortholog-pairs compared and the associated upper and lower stochastic bounds. The analysis performed refers to the following pairs of test species: Mouse *versus* Human, Medaka *versus* Human and Medaka *versus* Zebrafish. In essence, the proposed approach is new and computationally feasible in finding the maximum plausible extent of shared-synteny of the test species; and, it is found consistent with the underlying stochastic basis of Shannon information, entropy-theoretics and Schur-convexity.

Key words: Chromosomal synteny, synteny correlation, conserved shared-synteny, statistical association, Oxford-grid, Shannon's entropy, persisting mutual-information, Schur-convexity

INTRODUCTION

In general, the organisms of relatively recent divergence may show similar blocks of genes in the same relative positions in the genome. Known as the "synteny", [1 - 2], this situation corresponds to a block representing a set of contiguous genes located within the same chromosome; or, such a block is seen as a conserved entity between a pair of species or among various species, depicting a shared-synteny.

In practice, the extent of such synteny details is ascertained from a search-space pertinent to a limited set of samples of orthologous genomes of a given pair of (test) species. However, Houseworth and Postlethwait in [1] have suggested schemes to estimate the "true" number of conserved syntenies between the two test species by duly accounting the dependency of the number of orthologs searched towards chromosome-pairing. Relevantly, the present study is proposed to evolve, yet an enhanced and a computationally feasible strategy that accommodates the dependency of the number of orthologs involved when a large number (with a hypothetical limit tending to infinity) of orthologs are searched towards chromosome-pairing. Hence, it is attempted to find an answer to a relevant query posed by Postlethwait as indicated in [2], which can be stated as follows: "When we look at maps of conserved synteny between two species, we see that the orthologous genes are not randomly scattered

*Corresponding author:
Email: neelakan@fau.edu

http://dx.doi.org/10.20530/EJB_3_1-9
ISSN 2056-9912 © 2016

between the pairs of chromosomes. Some chromosome pairs contain many orthologs and some contain none. If we have mapped some (n) of the orthologs between two species and observe k conserved syntenies, can we estimate the number of conserved syntenies, ? , that will exist after we have mapped all the orthologs shared by both species?"

In answering the above query, the underlying approach involves the following: (i) Knowing the chromosomal organization *vis-à-vis* the synteny details available in a mapped format (such as, the so-called Oxford-grid [3 - 4] described in the next section); (ii) assessing the stochastic complexity and informational details (in Shannon's sense) [5] of the Oxford-grid; (iii) evolving entropy heuristic algorithms [6 – 7] to compute maximumally plausible extent of genetic similarity between the orthologous genomes in question; and lastly, (iv) applying the proposed methodology to find a solution for the Postlethwait problem concerning the following exemplary set of (test) species: Mouse (*Mus musculus*) *versus* Human (*Homeo sapiens*), Medaka (*Oryzias latipes*) *versus* Human (*Homeo sapiens*) and Medaka (*Oryzias latipes*) *versus* Zebra fish (*Donio rerio*).

Genomic Similarity AND SYNTENY CORRELATION: AN OVERview

Considering synteny conservation between two species (that is, the shared-synteny), the associated details can be represented as a matrix or mapping, framed by 'r' rows of chromosomes of one species *versus* and 'c' columns of chromosomes of the other species as illustrated in figure 1 where, the matrix [r × c] summarizes the map of the chromosome set $\{x_i\}$, (with i = 1 through c) of the first species X and the chromosome set $\{y_j\}$ (with j = 1 through r) of the second species Y. Correspondingly, suppose the genes on each of the chromosomes (i: 1 through r) are concurrently present in the chromosomes (j: 1 through c) and specified in the matrix cells implying the existence of of shared-synteny. Hence, the matrix framed with a total number of cells, m = (r × c) as shown in figure 1, commonly known as the Oxford grid [3 - 4], depicts the details on the state of homologous loci deduced from comparing the chromosomes of two species, X and Y.

Relevant to the analyses being pursued here, illustrated (partially) in figure 2 is an exemplar Oxford-grid availed from [8] for an orthologous pair of (test) species, namely, X: Mouse (*Mus musculus*) and Y: Human (*Homo sapiens*). Relevant data indicates that, there are [i = 1, 2, …, c = 19, vertical columns] and [j = 1, 2, …, r = 22, horizontal rows] so that, the total number of Oxford-grid cells is equal to m = (r × c) = 418 exhibiting a random dispersion of k (out of m) cells each showing a finite value of counts on conserved synteny. Correspondingly, there are (m – k) cells, each registering a null value on the synteny count. Full details of figure 2 are available, for example, in figure 12.2 of [8].

POSTLETHWAIT'S PROBLEM AND THE PROPOSED SOLUTION

In finding a solution on estimating the maximum plausible conserved synteny between species pairs observable in an exhaustive search-space of orthologous genomes, consistent with the query posed by Postlethwait [2], the step-by-step algorithmic pursuit involved is summarized below:

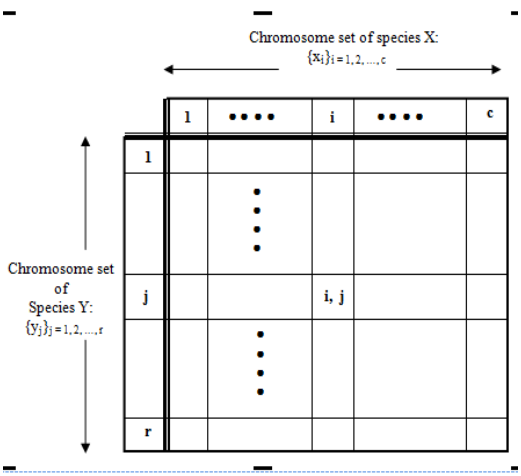

Fig 1: A matrix of the chromosome set $\{x_i\}$ of the species X and chromosome set $\{y_j\}$ of species Y.

Step I: Reference to the limited sample-space of n orthologous genomes searched pertinent to a test species-pair and the Oxford grid illustrated in figures 1 and 2, the associated parameters are summarized in Table 1; and, the underlying stochastic details imply probabilistic norms that associate the entropy (H) (or, persisting mutual-information (PMI) in Shannon's sense) with the data.

Table 1: Parameters of the Oxford-grid mapping (specified for a given pair of test species) depicting the number of cells with finite and null synteny counts relevant to: (A) Limited sample-space of orthologs described for example, in [8] [11]; and, (B) details on the new search-space with an infinitely large extent of orthologous genomes warranted in the context of solving the Postlethwait's problem [2]

(A) Known (limited) sample-space statistics available on the existing Oxford-grid pertinent to a pair of test species	Given a pair of test species, parameters of Oxford-grid mapping containing a total number of cells on the matrix: m = (r × c)		
	Number of orthologs pairs searched in a limited search-space	Number of cells showing finite synteny counts	Number of cells showing null synteny count
	n	k	(m − k)
(B) Unknown statistics to be deduced specific to an exhaustive search-space	For the given a pair of test species, the extended parameters of Oxford-grid mapping required in the context of solving the Postlethwait's problem		
	Number of ortholog pairs searched in an extended phase of exhaustive search-space	Resulting number of cells showing finite synteny counts in the new search-space	Corresponding number of cells showing null synteny count
	$n \rightarrow n_{uk} \rightarrow (N_{max} \rightarrow \infty)$	ℓ = ?	(m − ℓ) = ?

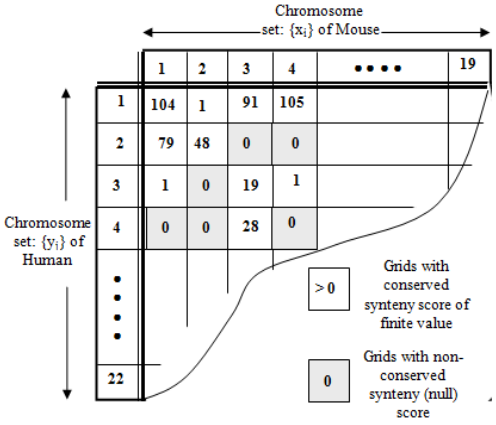

Fig 2: Oxford-grid layout (partial) with entry of synteny scores in the grids: An example for the test species Mouse *versus* Human. (Complete Oxford-grid is available in [8])

Step II: Reference to Table 1, the solution being sought on the Postlethwait's query, involves ascertaining the new value for (k ≤ m) → ℓ when n is increased to a large extent, (n → N_{max} → ∞). Correspondingly, the entropy details (H) pertinent to the random dispersion of (k ≤ m) cells in the test Oxford-grid will attain a maximum value, H_{max}. That is, when the search-space entity (n → n_{uk}) is monotonically increased, the trend of the associated entropy (of the Oxford-grid complex) will be, (H → H_{max}), leading to ascertaining the eventual extent of (k → ?) under the constraint that the total number of cells m = (r × c) remains invariant.

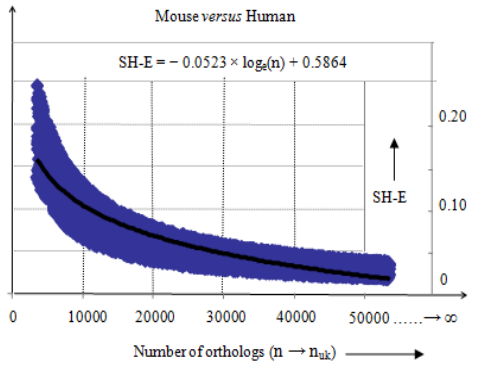

Fig. 3 Shannon entropy (SH-E) *versus* number of orthologous genomes (n → n_{uk}) adopted in the search-space of the test species, Mouse *versus* Human. SH-E is estimated using relevant Oxford-grid data of [8] and equation (6).

Relevant analytical details and computational steps are outlined in the following section.

ANALYTICAL PROCEDURE AND COMPUTATION

The analytical pursuit and computational details pertinent to the steps as above can be specified in a nut-shell in terms of the parameters of Table 1. It involves the procedural effort in tracking the sparse entropy, H towards its exhaustiveness specified by H_{max} such that, {H: n; k, (m – k)} → {H_{max}: (N_{max} > n_{uk}); (? ? k), (m – ?)}as outlined below:

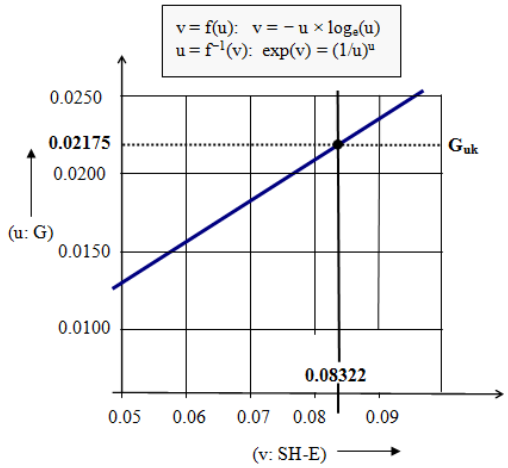

Fig. 4 Graphical solution to find the G-value (u) for a known entropy, SH-E (v) using the inverse entropy relation of [v = f(u) = − u × log$_e$(u)], namely, [u = f^{-1}(v): exp(v) = (1/u)^u]

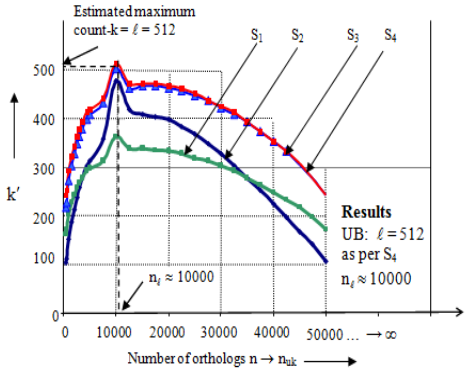

Fig. 5(b): Plot of simulation results on the variable k′ (denoting the synteny count-k) *versus* number of orthologs, (n_{uk} → N_{max} = 50,000) searched. The variable k′, (0 ≤ (k′ → ⍰) ≤ m denotes the number of cells on the Oxford-grid having a finite value of observed syntenic correlation at random cells, (i, j): Deduced for the Oxford-grid of the test species pair [11]: Medaka (*Oryzias latipes*) *versus* Human (*Homeo sapiens*)

1. Compilation of the data from the Oxford grid

Considering the example of test species (namely, Mouse *versus* Human), relevant Oxford-grid details in [8] show that the total number of orthologous genomes searched is: n = 3512. Other pertinent data observed in that test Oxford-grid are: (k = 154) denoting the number of cells having finite number of conserved-synteny scores marked out of a total number of m (= r × c) = (22 × 19) = 418 cells.

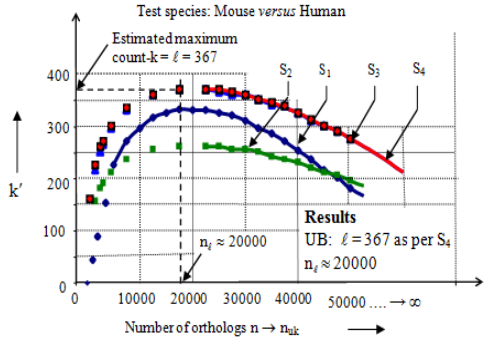

Fig. 5(a): Plot of simulation results on the variable k′ (denoting the synteny count-k) *versus* number of orthologs, (n_{uk} → N_{max} = 50,000) searched. The variable k′, (0 ≤ (k′ → ⍰) ≤ m denotes the number of cells on the Oxford-grid having a finite value of observed syntenic correlation at random cells, (i, j): Deduced for the Oxford-grid of the test species pair [8]: Mouse (*Mus musculus*) *versus* Human (*Homeo sapiens*)

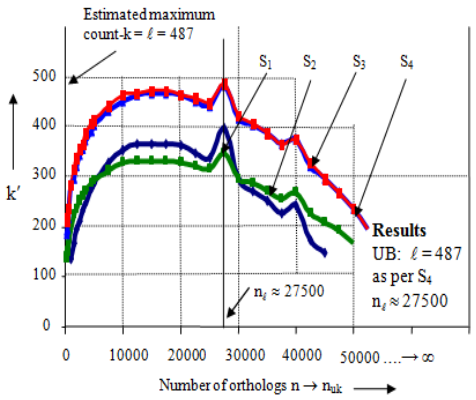

Fig. 5(c) Plot of simulation results on the variable k′ (denoting the synteny count-k) *versus* number of orthologs, (n_{uk} → N_{max} = 50,000) searched. The variable k′, (0 ≤ (k′ → ⍰) ≤ m denotes the number of cells on the Oxford-grid having a finite value of observed syntenic correlation at random cells, (i, j): Deduced for the Oxford-grid of the test species pair [11]: Medaka (*Oryzias latipes*) *versus* Zebra fish (*Donio rerio*)

Relevantly, the following ratios are first defined and calculated:

Ratio of the number of cells marked with a finite conserved synteny score to the total number of orthologous pairs searched:

D = (k /n) = (154/3512) (1)

Ratio of the number of cells marked with a finite conserved synteny score to the total number of cells in the Oxford-grid:

Table 2 Computation of: $\hat{k}_{S_1}$ (< m) using the metric: S_1

Data: Test Oxford-grid of Mouse *versus* Human [8] Search-space of orthologs: n_{uk} = 15,000 Elucidated Shannon entropy: G_{uk} $S_1 = [(k'/n_{uk})^{(k'/m)}] \times [(k'/n_{uk})^{(1-k'/m)}]$ with $(k' = 154) \leq \hat{k}_{S_1} \leq (m = 418)$			
k'	**S_1 Equation (8a)**	**G_{uk}**	**Maximum count-k determined *via* S_1 metric: $\hat{k}_{S_1} = k'$ at $(S_1 \approx G_{uk})$**
.			
284	0.0189333		
289	0.0192666		
294	0.0196000		
299	0.0199333		
304	**0.0202666**		Average of:
		0.02175	(304 and
309	**0.0206000**		309) ≈ **307**
314	0.0209333		
319	0.0212666		
324	0.0216000		
.			

$$E = [k/(m = r \times c)] = (154/418) \quad (2)$$

Ratio of the number of cells marked with a null score (implying non-conserved synteny status) to the total number of cells in the Oxford-grid:

$$F = [(m - k)/m] = [(418 - 154)/418] \quad (3)$$

2. Statistical aspects of randomly mixed grids as seen in Oxford-grid mapping

The Oxford-grid framework can be regarded as a receptacle dispersed with weighted proportions of randomly mixed entities, namely, k and (m – k) cells; and, this mixed state of such a binary set of cells depicts a universe of random space holding an entropy profile of uncertainty. Given such a random binary mixture, its effective probabilistic attribute can be quantitatively specified in terms of the so-called Lichtenecker-Rother (LR) formulation [9]; and, in the present context, this LR-formulation can be applied to the stochastic of the entities namely, the random-mix of the ratios, E = k/m and F = (m − k)/m. That is, the effective statistics of the mixed-states of k and (m – k) can be specified by a parametric entity G, which

Table 3 Computation of: $\hat{k}_{S_2}$ (< m) using the metric: S_2

Data: Test Oxford-grid of Mouse *versus* Human [8] Search-space of orthologs: n_{uk} = 15,000 Elucidated Shannon entropy: G_{uk} $S_2 = 1 - [(1 - k'/n_{uk})^{(k'/m)} \times (1 - k'/m)^{(k'/m)}]$ with $(k' = 154) \leq \hat{k}_{S_1} \leq (m = 418)$			
k'	**S_2 Equation (8b)**	**G_{uk}**	**Maximum count-k determined *via* S_2 metric: $\hat{k}_{S_2} = k'$ at $(S_2 \approx G_{uk})$**
.			
239	0.018199		
244	0.018965		
249	0.019745		
254	0.020542		
259	**0.021353**		Average of
		0.02175	(259 and
264	**0.022180**		264):≈ **262**
269	0.023022		
274	0.018267		
279	0.018600		
.			

can be deduced explicitly in terms of the set, {D, E, F} as follows:

$$G = D^E \times D^F = D^{(E+F)} \quad (4a)$$

Or,

$$\log_e(G) = E \times \log_e(D) + F \times \log_e(D) = (E + F) \times \log_e(D) \quad (4b)$$

Known also as the logarithmic law of mixing, the LR-formulation of equation (4) provides the true and effective description of the randomness of the underlying statistical mixture. However, the value of G evaluated as above, is constrained by the following Wiener inequalities [15, 16]:

$$[(D/E) + (D/F)]^{-1} \leq (G = D^E \times D^F) \leq [(D \times E) + (D \times F)] \quad (5)$$

The parameter G depicts a statistically justifiable value of effective probability ascribed to the randomness of the stochastic framework constituted by the set {D, E, F}; and, the underlying entropy details (in Shannon's sense) of this stochastic framework can be deduced as described below.

Table 4 Computation of: ⍰ ⍋$_3$ (< m) using the metric: S_3

Data: Test Oxford-grid of Mouse *versus* Human [8]
Search-space of orthologs: n_{uk} = 15,000
Elucidated Shannon entropy: G_{uk}
$S_3 = [(k'/n_1) \times (k'/m)]/[(1- k'/m) \times (k'/m)]$
with $(k' = 154) \leq$ ⍰ ⍋$_1 \leq (m = 418)$

k′	S_3 Equation (8c)	G_{uk}	Maximum count-k determined *via* S_3 metric: ⍰ ⍋$_3$ = k′ at ($S_3 \approx G_{uk}$)
`1			
344	0.019236421		
349	0.019810842		
354	0.020394214		
359	0.020986566		
364	**0.021587929**		Average of (364 and
369	**0.022198333**	**0.02175**	369):≈ **367**
374	0.022817809		
379	0.023446390		
384	0.024084106		
.			

Table 5 Computation of: ⍰ ⍋$_4$ (< m) using the metric: S_4

Data: Test Oxford-grid of Mouse *versus* Human [8]
Search-space of orthologs: n_{uk} = 15,000
Elucidated Shannon entropy: G_{uk}
$S_4 = [(k'/n_1) \times (k'/m)]$
with $(k' = 154) \leq$ ⍰ ⍋$_1 \leq (m = 418)$

k′	S_3 Equation (8d)	G_{uk}	Maximum count-k determined *via* S_4 metric: ⍰ ⍋$_4$ = k′ at ($S_4 \approx G_{uk}$)
.			
349	0.019425997		
354	0.019986603		
359	0.020555183		
364	0.021131738		
369	**0.021716268**		Average of
		0.02175	(364 and
374	**0.022308772**		369):≈ **367**
379	0.022909250		
384	0.0235177030		
389	0.0241341310		
.			

3. Entropy of randomly mixed finite and null score grids in the Oxford grid mapping

The Shannon entropy of G (denoted as SH-E) can be written as follows [5]:

$$\text{SH-E} = -\, G \log_e (G) \quad \text{nats} \quad (6)$$

Equation (6) represents the uncertainty framework of the test Oxford-grid containing sparse synteny data availed from a limited search exercised on synteny correlation of the test pair of species as specified by the set, {n; k, (m – k)} of Table 1. This sparse entropy status can be extended to decide on the maximum plausible value (H_{max}) of the resulting entropy functional pertinent to the set {⍰, (m – ⍰)} that would eventually prevail in the Oxford-grid, when $n \rightarrow (N_{max} \rightarrow \infty)$, subject the constraint stated as, $(k \leq ℓ \leq m)$.

Towards computing this new value of entropy (H_{max}), first, the extent of search-space of the orthologous pairs, namely, n is set as a variable, n_{uk} and it is rendered to increase say, linearly, in the range, $3512 \leq n_{uk} \leq (50{,}000 \rightarrow \infty)$. Next, in order to determine the unknown value of $(k \rightarrow k')$ for each value of $(n \rightarrow n_{uk})$, a set of uniformly-distributed random numbers can be assumed for the unknown k′ ⍰within the constraining limits, $(k = 154) \leq k' \leq m (= 418)$ commensurate with the Laplacian hypothesis of unknown statistics. Hence, the ratios defined in equation (4) can be calculated for each set $\{n_{uk}; k', m\}$. The associated value of G and the corresponding estimate of entropy (SH-E) can then be determined *via* equations (4) and (6) respectively. Thus, the resulting entropy, SH-E *versus* $n \rightarrow [n_{uk} \rightarrow (N_{max})]$ is plotted as illustrated in figure 3; here, the patch of results seen along the ordinate denotes the results obtained with an ensemble of randomly-chosen (uniformly-distributed) set of values assigned to k′ such that , $(k = 154) \leq k' \leq m (= 418)$, for each, linearly incremented value of n_{uk} in the range, $(3512 \leq n_{uk} \leq 50{,}000)$.

Table 6: Computed data and a summary on results relevant to Postlethwait's query [2]

Test species pairs	Available Oxford-grid data of the test species pairs: {**k, n**}, **m** and **(m – k)**}			Estimated values: (by the present method) **k →** ($\ell \leq$ **m)** at n_{ℓ} when n → (N_{max} = 50000) keeping the value of m constant as specified by the Oxford-grid				
				Proposed metrics				At **(n → n_{ℓ})** ≈
					LB		**UB**	
				S_1	**S_2**	**S_3**	**S_4**	
	{k, n}	**m** (r × c)	**(m – k)**	Estimated **k →** ($\ell \leq$ **m)**				
				ℓ_{S1}	ℓ_{S2}	ℓ_{S3}	ℓ_{S4}	
Mouse *versus* Human [8]	**{154, 3512}**	**418** (22 × 19)	**304**	307	262	367	**369**	**20000**
Medaka *versus* Human [11]	**{104, 818}**	**552** (24 × 23)	**448**	477	362	502	**512**	**10000**
Medaka *versus* Zebrafish [11]	**{125, 255}**	**600** (24 × 25)	**475**	397	347	483	**487**	**27500**

Summary of the results

Available sparse Oxford-grid data of the test species pairs obtained from a limited synteny search and specified in terms of the following parameters: **[k** and **n]**	Answers to Postlethwait's query on maximum plausible shared synteny: Estimated values obtained in an exhaustive search-space of (n) → (n_{uk} → N_{max} → ∞) using the proposed method and expressed in terms of the following parameters: **[(k →** ℓ**)** at n_{ℓ} **]**
Mouse *versus* Human [8] **[154** and **3512]**	**[369** at **20000]**
Medaka *versus* Human [11] **[104** and **818]**	**[512** at **10000]**
Medaka *versus* Zebrafish [11] **[125** and **255]**	**[487** at **27500]**

Now, for the clustered patch of computed details as seen in figure 3, a regressed, isotonic trend-line is obtained with coefficients (of regression), α_1 and α_2 and expressed as follows:

$$SH\text{-}E[(n \rightarrow n_{uk})] = -\alpha_1 \times \log_e [(n \rightarrow n_{uk})] + \alpha_2 \quad (7)$$

For the test-species (Mouse *versus* Human) being considered, this trend-line so estimated has the values, $\alpha_1 = -0.0523$ and $\alpha_2 = 0.5846$ (in figure 3). Hence, for any presumed (and unknown) value of (n → n_{uk}) searched in the range, 3512 ≤ n_{uk} ≤ (50,000 → ∞), the associated entropy (SH-E) can be estimated with equation (7). For example, relevant computation yields, SH/E = 0.08322 for an assumed value of (n → n_{uk}) = 15000.

4. Determining the possible value of ℓ > (k ≤ m) *via* an expanded search on orthologs

The next step involves finding the value of G corresponding to a known (estimated) entropy value SH-E. For example, considering SH-E (for n_{uk}= 15000) = 0.08322 as assessed above, an inverse solution of the entropy relation, SH-E(n_{uk}) = − G_{uk}× $\log_e(G_{uk})$ of equation (6) can be adopted to find the corresponding value of G_{uk}. That is, considering the entropy equation in its general form: v = f(u) = − u $\log_e$ (u), relevant inverse relation is given by u = f^{-1}(v); or, explicitly, $(1/u)^u$ = exp(v), with v: SH-E and u: G. However, the relation, $(1/u)^u$ = exp(v) is a transcendental equation and as such, a simple analytical solution for u for a given value of v, is not feasible. But, a graphical solution can be attempted as illustrated in figure 4, where the entities (v: SH-E) *versus* (u: G) are plotted with u incremented linearly from an initial (small) value (say, for example 0.0001, 0.0002 …etc.); and, the corresponding Shannon entropy, namely, v: SH-E is computed and plotted as a dependent variable as shown. Now, a value G_{uk}

is identified and marked on the u *versus* v graph such that, it corresponds to the intersection of the computed, (u: G) *versus* (v: SH-E) curve and the known value of v: SH/E = 0.08322 (corresponding to, n_{uk} = 15,000) deduced earlier. Thus, the solution for the G-value being sought (and denoted as G_{uk}) is found to be 0.02175 as shown in figure 4. Thus, for every value of n_{uk} in the range, ($3512 \leq (n \rightarrow n_{uk}) \leq 50{,}000$), corresponding value of G_{uk} can be determined.

Using the deduced value of G_{uk} as above, the procedure towards elucidating $k' \rightarrow (\ell \leq m)$, with the ortholog search-space of $n_{uk} \rightarrow N_{max}$ can be attempted as follows: The variable k′ is increased (say, in the increments of 5), from its base value to the maximum value, m = (r × c). In each case of k′ so chosen, the Oxford-grid denotes a domain with a new stochastic state invoked by the presence of a random mix of entities, namely, k′ and (m − k′) cells. Hence a set of metrics implicitly depicting the G-value of LR-heuristics can be defined, along with the statistical limits of upper- and lower-bounds (LB and UB) as follows:

Statistical random mixture LR-formulation based metric [9]:

$$S_1 = [(k'/n_{uk})^{(k'/m)}] \times [(k'/n_{uk})^{(1-k'/m)}] \quad (8a)$$

Statistical susceptance metric yielding the lower-bound (LB) (Neelakanta's susceptibility formula [10])

$$S_2 = 1 - [(1- k'/n_{uk})^{(k'/m)} \times (1- k'/m)^{(k'/m)}] \quad (8b)$$

Statistical Wiener upper-bound (UB) metric (version 1)

$$S_3 = [(k'/n_1) \times (k'/m)]/[(1- k'/m) \times (k'/m)] \quad (8c)$$

Statistical Wiener upper-bound (UB) metric (version 2, known as Beer measure [10]):

$$S_4 = [(k'/n_1) \times (k'/m)] \quad (8d)$$

As stated above, the metrics specified by equations (8a)–(8d) denote the variates of LR-formulation denoting the G-value consistent with stochastic mixture theoretics. Specifically, the metric S_1 of equation (8a) exactly refers to the LR-formulation of equation(4); and, the other metric, namely, S_2, (S_3 and S_4) respectively denote the lower- and upper-bounds of S_1 as per the Wiener-limits.

For a given set of {n; k, (m – k)} and a chosen value of $n \rightarrow n_{uk}$, the metrics S_1, S_2, S_3 and S_4 can be computed, each as a function of the variable, $k \rightarrow k'$. In each case, the particular value of k′ that yields the assessment of the metric in question matching G_{uk} (deduced as explained earlier) would denote the maximum count-k, namely, □ (≤ m) being sought for the search-space of n_{uk}.

For example, as presented in the Table 2, considering the data set{n_{uk} =15000, G_{uk} = 0.02175}, the assessed variation of S_1 *versus* k′ shows the value of the count-k ascertained with the metric, $S_1 = G_{uk} =$ 0.02175 as □ equal to: (304 + 309)/2 = 307; that is, this value of □ depicts the maximum possible extent of cells in the Oxford-grid mapping that would show a finite extent of conserved synteny, enumerated in the search-space with n_{uk} = 15,000 orthologous genomes of the test species. *In lieu* of the metric S_1, other measures namely, S_2, S_3 and S_4 can also be prescribed in the procedure narrated above. Corresponding results of the computations are denoted respectively as: □, □ and □ and presented in Tables 3 –5, using the same data set adopted in the case of □ assessment with the metric: S_1.

COMPUTED RESULTS

Relevant to the Oxford-grid data-set of the test species, stretching the parameter $n \rightarrow n_{uk}$ to range from a low value, say, 500 to N_{max} = 50,000 ($\rightarrow \infty$), the maximum plausible count-k is obtained for each value of n_{uk} and expressed in terms of the set {□, □, □, □} as indicated in figures 5(a) - 5(c), depicting the plots of k′ *versus* ($n \rightarrow n_{uk}$) pertinent to the following test species respectively: (i) Mouse *versus* Human [8], (ii) Medaka *versus* Human [11] and (iii) Zebrafish *versus* Medaka [11].

In each case of the species pairs considered, relevant graphs in figures 5(a) - 5(c) show an initial monotonic increasing trend of k′ *versus* $n \rightarrow n_{uk}$; and then, the plots assume a convex form with a maximum value of $k' \rightarrow$ □ ata certain value of $n_{uk} = n_{□□}$. The details in figures 5(a) - 5(c) also include the associated statistical lower- and upper-bounds on S_1 as decided respectively by the Wiener limits, S_2 and (S_3, S_4). In each case of the test species pairs studied, a

summary of the pertinent results is presented in Table 6.

DISCUSSION

In seeking an answer on the maximum plausible extent of syntenic correlation that would prevail between the chromosomes of a pair of test species, when their orthologous genomes are searched exhaustively, the proposed strategy hypothesizes that, the limited statistics of k (out of m entities) in the Oxford-grid would dictate a sparse value of underlying (Shannon's) entropy. Such entropy features stem from the statistically-mixed pair of cell entities, k and (m – k) in the Oxford-grid and depict implicitly the information on the patterns of weighted probability that describe the effective stochastic attributes of the associated mixture constituents, as indicated by Neelakanta et al. in [12]. Correspondingly, the proposed metric S_1 measures the effective statistical attributes (of the mixture constituents in the Oxford-grid); and, it is bounded by the statistical Wiener limits, given by: $[S_2 < S_1 < (S_3, S_4)]$.

Further, the assessed sparse-entropy plots in figures 5(a) – 5(c) show a convexity profile, when the search is extended on more and more number of pair of orthologs with $n \rightarrow (n_{uk} \rightarrow N_{max} \rightarrow \infty)$. Relevant simulated results conform to Schur-convex functional attributes consistent with maximum entropy heuristics [6].

CONCLUDING REMARKS

In all, this study is a focused effort in finding an answer for the Postlethwait's query concerning syntenic correlation attributes *vis-à-vis* the posterior (*ex ante*) status of shared synteny that can be inferred in an exhaustive search made on test orthologous pairs, by projecting the sparse data (availed *ex post* from the test Oxford-grid). The proposed methodology refers to stochastic modeling of the associated Oxford-grid parameters and applying entropy-theoretic considerations. To the best of authors' knowledge no such prior study exists on the relevant estimation. Efficacy of the proposed approach is implicitly confirmed by the Schur-convexity of entropy–based results obtained in conformance with the analytics of maximum entropy optimization principles [6].

REFERENCES

1. Housworth EA, Postlethwait J. Measures of synteny conservation between species pairs. Genetics. 2002 Sep;162(1):441-8. PubMed PMID: 12242252; PubMed Central PMCID: PMC1462247.
2. Houseworth, E. A,, 2003. Measure of Conserved Synteny. Abstract for IMA: RECOMB Satellite Workshop on Comparative Genomics (Oct. 20-24, 2003, University of Minnesota). Available at: http://www.ima.umn.edu/2003-2004/W10.20-24.03/activities/Housworth-Elizabeth/housworth.pdf.
3. EDWARDS JH. The Oxford Grid. Ann Human Genet [Internet]. Wiley-Blackwell; 1991 Jan;55(1):17–31. Available from: http://dx.doi.org/10.1111/j.1469-1809.1991.tb00394.x
4. OXGRID–The Oxford Grid Project. Available at: http:// oxgrid.angis.org.au/
5. Neelakanta, P. S., 1999. Information-Theoretic Aspects of Neural Networks, CRC Press.
6. Kapur JN, Kesavan HK. Entropy Optimization Principles and Their Applications. Water Science and Technology Library. 1992;3–20. Available from: http://dx.doi.org/10.1007/978-94-011-2430-0_1.
7. Yee J, Kwon M-S, Park T, Park M. A Modified Entropy-Based Approach for Identifying Gene-Gene Interactions in Case-Control Study. Li Y, editor. PLoS ; 2013 Jul 18;8(7):e69321. Available from: http://dx.doi.org/10.1371/journal.pone.0069321
8. Pearce DA. Microarray Analysis. Mark Schena. New York: Wiley-Liss, John Wiley & Sons, Inc., 2003, 654 pp., $89.95, hardcover. ISBN 0-471-41443-3. Clinical Chemistry. 2003 Jun 1;49(6):1031–1031. Available from: http://dx.doi.org/10.1373/49.6.1031
9. Lichtenecker K. and Rother K. Die Herleitung des logarithmischen Mischungsgesetzes aus allgemeinen Prinzipien der statioenaeren Stroemung. Physikalische Zeitschrift. 1938; 32: 255-260.
10. Neelakanta, P. S., 1995. Handbook of Electromagnetic Materials. CRC Press, pp.166.
11. Naruse K. A Medaka Gene Map: The Trace of Ancestral Vertebrate Proto-Chromosomes Revealed by Comparative Gene Mapping. Genome Research. 2004 May 1;14(5):820–8. Available from: http://dx.doi.org/10.1101/gr.2004004
12. Neelakanta PS, Arredondo TV, and De Groff D. Redundancy Attributes of a Complex System: Applications in Bioinformatics. Complex System. 2003; 14: 215-233.