BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 9(10)

Sequences encoding identical peptides for the analysis and manipulation of coding DNA

Joaquín Sánchez

Facultad de Medicina, UAEM, Calle Ixtaccihuatl Esq Leñeros, Col. Los Volcanes C.P. 62350, Cuernavaca, Morelos, Mexico; Joaquín Sánchez – Email: joaquin.sanchez@microbio.gu.se; Phone: 52-777-3184797; *Corresponding author

Received May 16, 2013; Accepted May 19, 2013; Published June 08, 2013

Abstract:
The use of sequences encoding identical peptides (SEIP) for the *in silico* analysis of coding DNA from different species has not been reported; the study of such sequences could directly reveal properties of coding DNA that are independent of peptide sequences. For practical purposes SEIP might also be manipulated for e.g. heterologous protein expression. We extracted 1,551 SEIP from human and *E. coli* and 2,631 SEIP from human and *D. melanogaster*. We then analyzed codon usage and intercodon dinucleotide tendencies and found differences in both, with more conspicuous disparities between human and *E. coli* than between human and *D. melanogaster*. We also briefly manipulated SEIP to find out if they could be used to create new coding sequences. We hence attempted replacement of human by *E. coli* codons via dicodon exchange but found that full replacement was not possible, this indicated robust species-specific dicodon tendencies. To test another form of codon replacement we isolated SEIP from human and the jellyfish green fluorescent protein (GFP) and we then re-constructed the GFP coding DNA with human tetra-peptide-coding sequences. Results provide proof-of-principle that SEIP may be used to reveal differences in the properties of coding DNA and to reconstruct in pieces a protein coding DNA with sequences from a different organism, the latter might be exploited in heterologous protein expression.

Keywords: synonymous codons, intercodon dinucleotides, codon pairs, codon allocation tendencies, green fluorescent protein, heterologous protein expression.

Background:

Due to the nature of the genetic code in human and in many other organisms, apart from two cases (methionine and tryptophan), more than one equivalent (synonymous) codon can be used by the cell to specify the same amino acid **[1]**. This versatility is expressed differently in organisms so that they vary in codon usage **[1-4]** and this, as well as the arrangement of synonymous codons, may affect protein translation **[5-8]**. Variations in codon usage between organisms are usually estimated in entire collections of protein-coding sequences (ORFeome). Studies in ORFeomes have led to the identification of context-dependent regularities in codon bias **[9]** and of tendencies in dicodon frequency **[10, 11]**. Dicodon tendencies naturally reflect in intercodon dinucleotide frequencies, which may be a strong selective force in genes **[12]**.

A different alternative to investigate codon usage and synonymous codon distribution is the analysis of sequences encoding identical peptides in separate organisms. In principle, such an approach avoids potential effects of amino acid composition **[13]** and protein sequence **[14]**. We thus explored if the comparison of sequences encoding the same peptides in different organisms could directly reveal specific tendencies in coding DNA. After initial practical examination of the proposed strategy, it became clear that in order to have a manageable amount of data, and at the same time a not too small number of sequences, the length of encoded peptides could not be less than 6 amino acids. We therefore collected sequences encoding peptides ≥6 amino acids in length from human, *Drosophila melanogaster* and *Escherichia coli*.

We evaluated tendencies in synonymous codon distribution through the comparison of intercodon dinucleotide frequencies in native and synonymous codon-shuffled sequences, as earlier described **[15].** With such method distinct propensities were found in human, in *D. melanogaster* and in *E. coli*. In addition, we briefly investigated the possibility of generating new sequences with converted codon usage for possible heterologous protein expression.

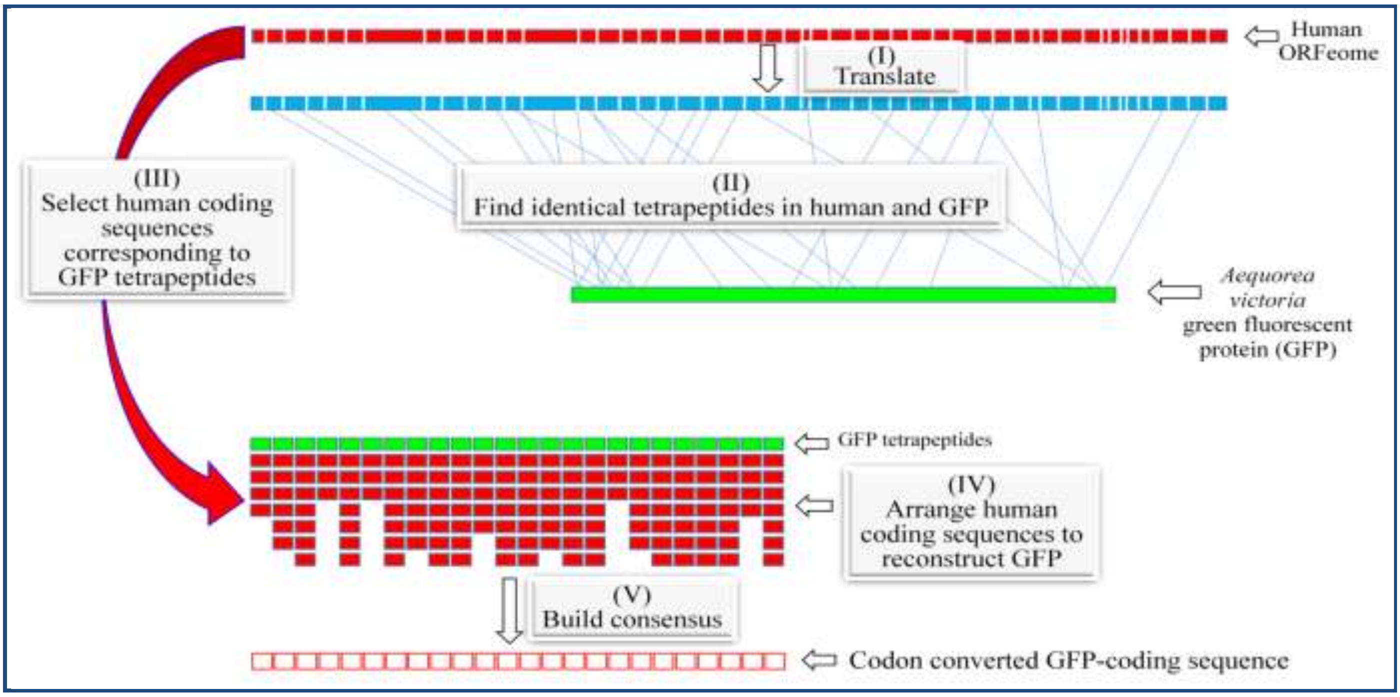
Human ORFeome
(I) Translate
(III) Select human coding sequences corresponding to GFP tetrapeptides
(II) Find identical tetrapeptides in human and GFP
Aequorea victoria green fluorescent protein (GFP)
GFP tetrapeptides
(IV) Arrange human coding sequences to reconstruct GFP
(V) Build consensus
Codon converted GFP-coding sequence

Figure 1: Scheme depicting the procedure to replace GFP codons by human codons using human sequences encoding identical tetrapeptides. Roman numerals in parentheses are used to indicate the sequence of the process. Drawings are not to scale. The horizontal bars represent either the human ORFeome (red bar on top) as indicated, or the collection of human proteins (blue) upon ORFeome translation (Roman numeral I). The green horizontal bar represents the GFP protein. As indicated, a segmented green horizontal bar is used to represent 59 tetrapeptides integrating GFP, except for the last two amino acids. The red squares below the segmented GFP green bar represent the multiple human sequences used to reconstruct the GFP coding sequence after defining a consensus (Roman numeral V) for each tetrapeptide-coding sequence.

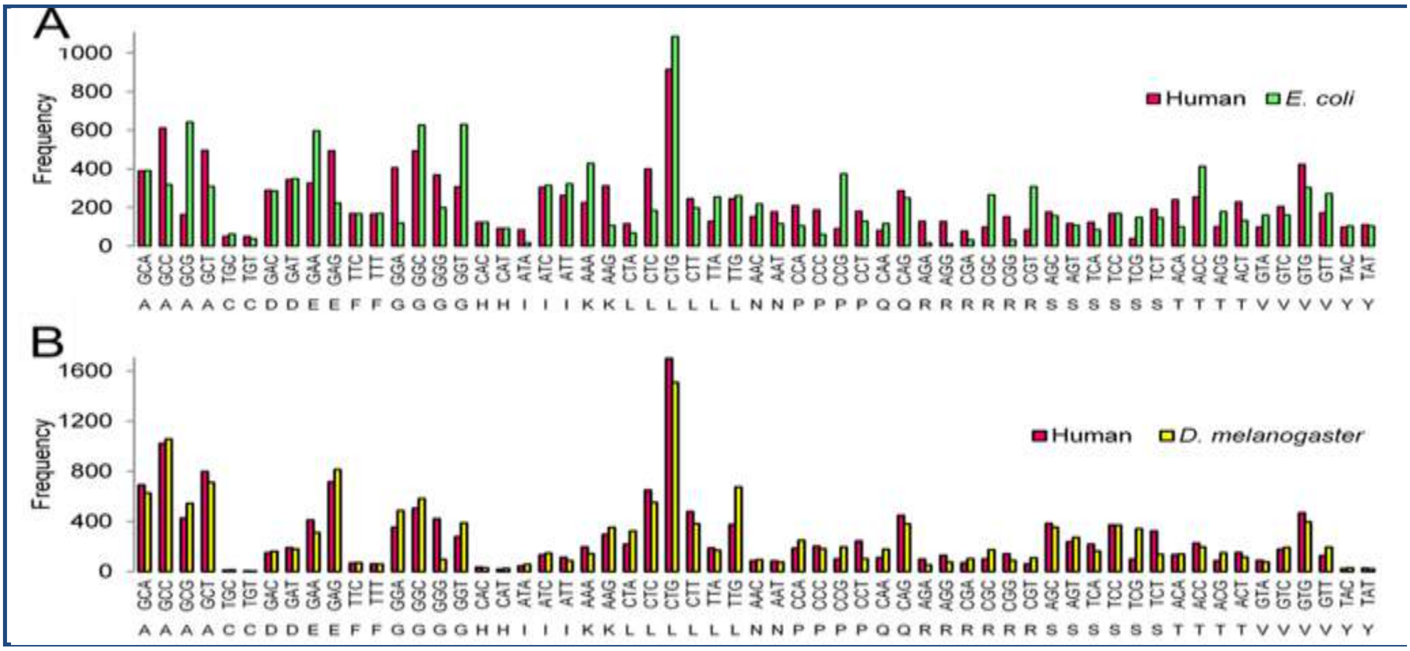

Figure 2: Codon frequencies in sets of sequences encoding identical peptides. Human sequences were different in **A** and **B.** The origin of codons is indicated above each graph. The x-axis shows codon sequences and the corresponding encoded amino acid (one letter code). In both panels neither methionine nor tryptophan codons are shown as they had identical frequency because they are encoded by a single codon each. In the y-axis absolute codon frequencies are shown.

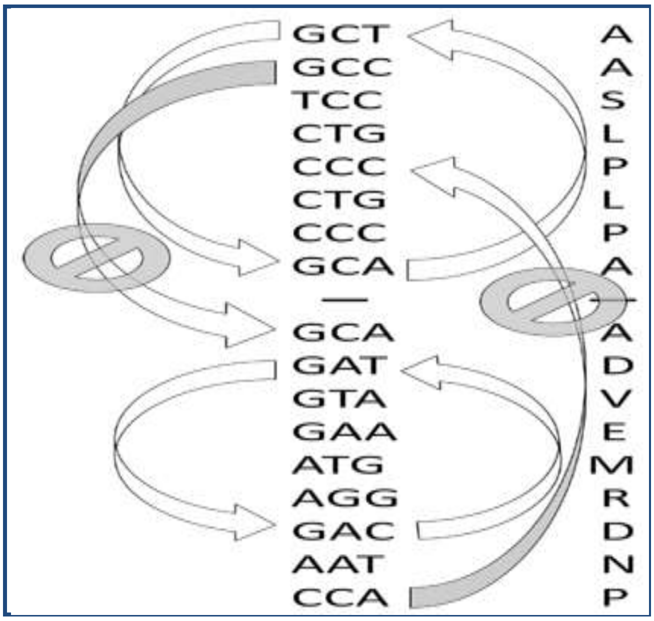

Figure 3: Procedure used to exchange synonymous codons in sequences. In the drawing the exchange of synonymous codons is represented. Arrows indicate that for each peptide-coding sequence the exchange of synonymous codons is allowed internally but not with sequences encoding other peptides (symbol forbidden over arrows).

Methodology:

Data retrieval and software

The human ORFeome version hORFeome v3.1 consisting of over 12,000 human coding sequences was downloaded from http://horfdb.dfci.harvard.edu/. Annotations were edited out and sequences were curated to remove out-of-frame sequences, sequences lacking stop codons, and sequences not starting with ATG. The enteropathogenic E2348/69 *Escherichia coli* (*E. coli*), *Drosophila melanogaster* and the *Aequorea victoria* green fluorescent protein coding sequences were obtained from http://ncbi.nlm.nih.gov and, if needed, they were equally curated before analysis. To collect and analyze sequences we used either one or a combination of the programs Word (Microsoft Inc, MS), Excel (MS) and OMIGA 2.0 (Oxford Molecular Ltd, UK).

Sequences encoding identical peptides

The above described coding sequences were translated *in silico* and the produced protein sequences were compared using the OMIGA Dot Plot program to identify equal peptides in the different organisms. The identified peptide sequences were then used to retrieve the partnering coding sequences with the OMIGA Dot Plot in the mode DNA vs. protein. The gathered coding sequences were then filtered in Excel to remove duplicates and to produce equally-sized collections of human and *E. coli* sequences (1,551 sequences, **Table 1 (Available with authors)** and human and *D. melanogaster* sequences (2,631 sequences, **Table 2 (Available with authors).** The final human *E. coli* set was composed of sequences encoding peptides 8-16 amino acids long while the human-*D. melanogaster* set was comprised of sequences encoding peptides 6 amino acids long.

Shuffling of synonymous codons

The shuffling of coding sequences was carried out in an Excel spreadsheet and the process was mainly based in the controlled ordering of cells in columns. Basically, individual codons and their matching amino acids were placed in adjacent columns, an extra column that contained identifying labels for each peptide-coding sequence was also inserted. For the purpose of shuffling an additional column with random numbers was used to rearrange synonymous codons without changing the ordering of amino acids; the identifying labels served to prevent exchange between sequences. Six sequential synonymous codon shuffles were carried out and five independent replicas were generated to compute the mean and standard deviation.

Intercodon dinucleotide frequencies

We calculated differences in intercodon dinucleotide frequencies as before **[15]** by comparing native frequencies to those in the shuffled sequences, except that we here report the percent increment, or decrement, in intercodon dinucleotide frequency.

Codon replacement

For replacement of human for *E. coli* codons the human sequences were shuffled together with *E. coli* sequences using a procedure analogous to the one described above. However, in this case codon identifying labels were arranged so that there was interchange between human and *E. coli* sequences if they coded for the same peptide. The shuffling mixture was composed of one copy of human sequences and 68 copies of the *E. coli* sequences. The use of excess copies of the *E. coli* sequences increased the probability of replacement of human codons for *E. coli* codons. Additionally, codon replacement involved the swapping of pairs of synonymous codons, which was done to preserve global dinucleotide composition. The method was designed so that it allowed only the exchange of codon pairs in which the individual codons differed at the wobble position, e.g. the AGCCTC (SerLeu) codon pair could be exchanged by AGTCTT (SerLeu) but not by TCCCTG (SerLeu) or viceversa.

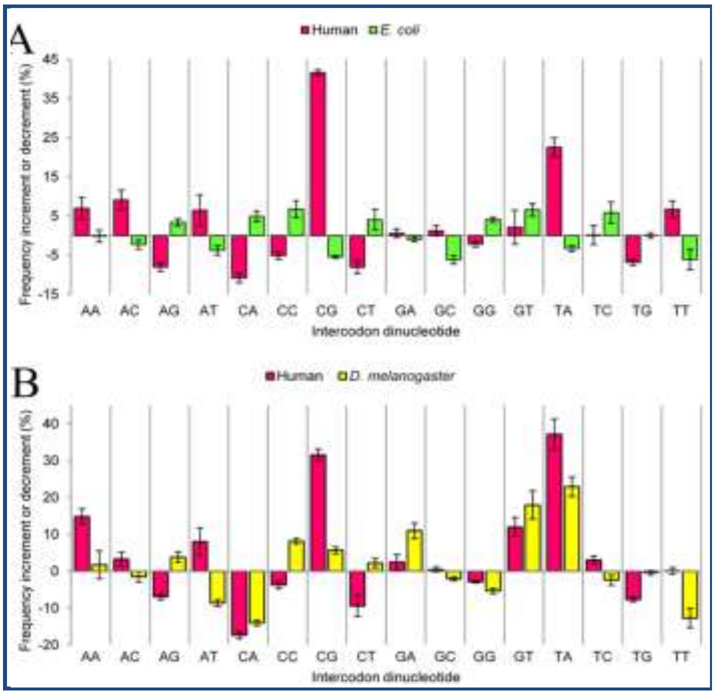

Figure 4: Percent increment or decrement in intercodon dinucleotide frequencies after shuffling of synonymous codons in sequences encoding identical peptides in human and *E. coli* (A) or in human and *D. melanogaster* (B). In the x-axis intercodon dinucleotide sequences are shown. In the y-axis the

change in percentage in intercodon dinucleotide frequency is shown. Standard deviations are shown above bars.

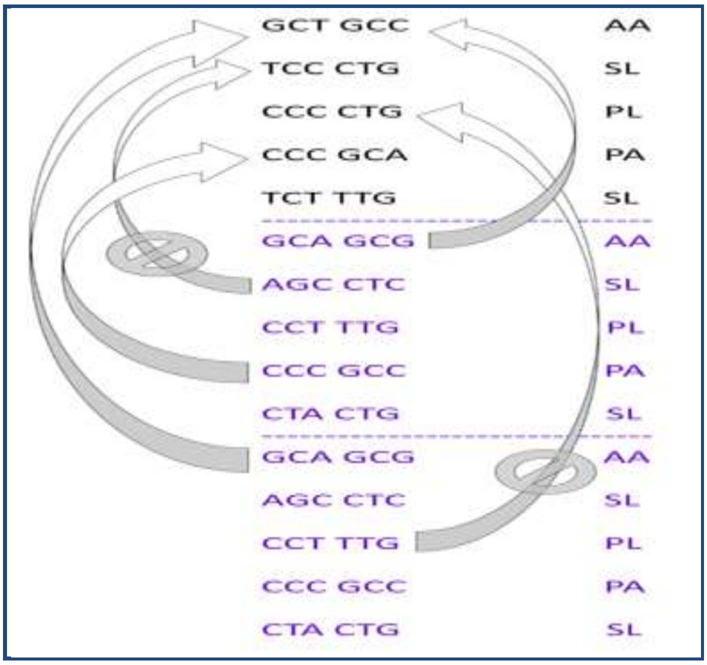

Figure 5: Procedure used to replace human for *E. coil* codons by exchange of pairs of synonymous codons in sequences coding for identical peptides. The exchange occurs between one copy of the human sequences (sequence on top) and 68 copies of sequences from *E. coli* (symbolized by two sequences in the bottom). The dotted line indicates the boundary between independent peptides. Arrows indicate the exchange of pairs of synonymous codons. The forbidden symbol over arrows indicates changes that are not allowed. Codon exchange downwards, i.e. towards *E. coli* sequences, is not shown because its effects are virtually irrelevant due to the disparity in the number of copies.

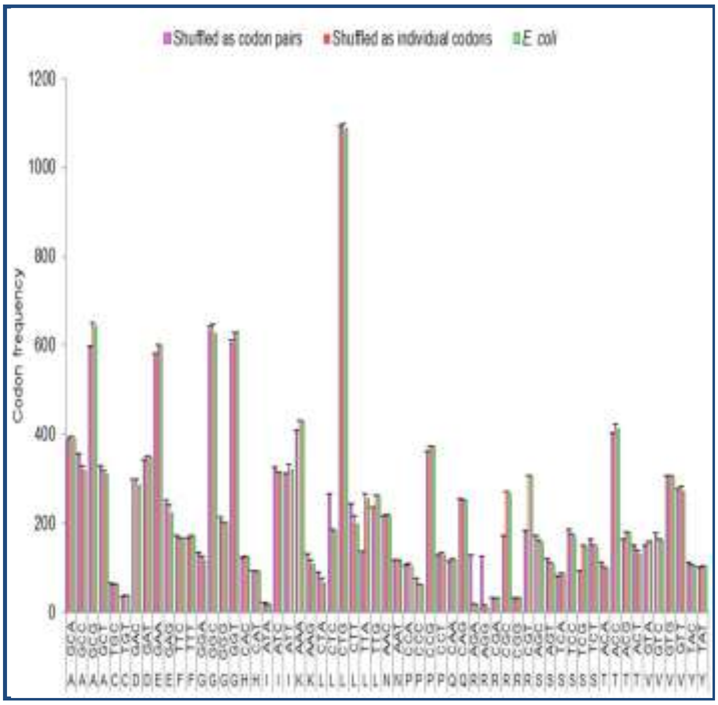

Figure 6: Codon composition of human sequences encoding identical peptides before and after shuffling either as pairs of synonymous codons or as individual codons for replacement of human codons by those in *E. coli*. In the x-axis codon sequence and encoded amino acid one letter code) are shown. The y-axis shows absolute frequency. Above bars of shuffled sequences standard deviations are shown. For comparison, codon compositions of intact *E. coli* sequences are shown.

Results & Discussion:

Search for identical peptides

The basis of our proposal is the identity between peptides from different organisms, so we investigated if such identical peptides could occur by chance. We hence randomly swapped amino acids in *E. coli* proteins prior to comparison with human proteins. After such shuffling virtually no equivalent peptides were detected, which indicated that peptide sequence identity was dependent on biological protein sequences. Actually, peptide sequence identity could also be demonstrated with standard data search procedures such as the BLAST program at http://www.ncbi.nlm.nih.gov provided compositional adjustments are not applied. In this way BLAST would promptly reveal numerous identical peptide sequences in human and *E. coli* that are often part of similar proteins.

Codon usage in sequences encoding identical peptides

To investigate if there were differences between human and *E. coli* sequences coding for the same peptides we determined codon frequencies, given that sequences encoded the same peptides this was equivalent to estimating codon usage. We present codon frequencies for 1,551 human and *E. coli* sequences coding for the same peptides **(Figure 2A)**. As seen, there were disparities in codon frequency between the two organisms among which perhaps the most noticeable are the higher frequencies of the GCG, CCG, CGC and CGT codons in *E. coli* than in human and the lower frequencies of the arginine codons AGA and AGG in *E. coli* than in human. In the first case, the higher relative frequency in *E. coli* could be explained by an overall avoidance of the dinucleotide CG in human DNA **[16-18]**, although in the present analysis CGG would be an exception to such proposal. Conversely, the lower relative frequency of AGA and AGG codons in *E. coli* possibly relates to potential interference with protein expression **[19]**. The reason for it may be, as proposed **[7]**, avoidance of latent complementarity with the anti-Shine-Dalgarno sequence in the 16S ribosomal RNA, which could slow down translation. If such proposition is correct, a higher occurrence of AGA and AGG codons in human sequences would be explained by the absence of an anti-Shine-Dalgarno sequence in human ribosomal RNA.

In the above example human sequences were compared to prokaryotic sequences, we then wished to determine if analogous results would be obtained if sequences were from less distant organisms. Codon frequency analysis for 2,631 sequences coding for the same peptides in human and *D. melanogaster* **(Figure 2B)** showed differences in codon usage that were not as marked as between human and *E. coli*, but noticeable dissimilarities are the predominance of the TCG serine codon in *D. melanogaster* and the reciprocal preponderance of the TCT serine codon in human. These differences could also be associated to the tendency to avoid the dinucleotide CG in human since there was a mild, but persistent, lower frequency of codons CCG, CGC, CGA and CGT in human sequences **(Figure 2B).**

Intercodon dinucleotide frequencies

The above examples demonstrated utility of the proposed approach to study codon usage differences between species, but besides codon usage there is another interesting property that should be measurable, especially under constant peptide sequence, i.e. tendencies in synonymous codon allocation. One way that kind of tendencies can be assessed is by comparison of intercodon dinucleotide frequencies. Intercodon dinucleotides are of interest because they can be a strong selective force in genes **[12]** and they have been proposed to affect 3-base periodicity in coding DNA **[15]**, which may in turn be important for gene expression **[20].**

To investigate differences in intercodon dinucleotides we contrasted native frequencies with frequencies in a shuffled control. Shuffling of synonymous codons in sequences was carried out as schematized **(Figure 3),** the scheme shows that synonymous codon exchange was allowed only internally, in this way codon composition of individual sequences was completely preserved.

To help appreciate the degree of variation in intercodon dinucleotide frequencies after shuffling of synonymous codons we calculated the frequency change in percentage. Percent changes in intercodon dinucleotide frequencies are shown **(Figure 4A)** for human and *E. coli*, among the most prominent dissimilarities are the increments in CG and TA in human. The result means that in native human sequences NNC codons (where N may be A, C, G or T) tend to avoid GNN codons as their immediate downstream neighbors while NNT codons tend to avoid ANN codons. This agrees with previously reported data [15]. Oppositely, in *E. coli* sequences there was no avoidance of CG or TA dinucleotides, in fact, even though with relatively small percentage values, results would rather suggest preference of codons NNC and NNT for codons GNN and ANN respectively in *E. coli*.

In contrast to the differences between human and *E. coli* sequences, where most changes in intercodon dinucleotide frequencies occurred in opposite directions, in human and *D. melanogaster* changes happened both in opposite directions and in the same direction **(Figure 4B).** Noticeably, dinucleotide frequency changes in the two sets of human sequences were qualitatively very similar, even though the sequences used for comparison with *E. coli* **(Figure 4A)**were different to those used for comparison with *D. melanogaster* **(Figure 4B).** This suggests that similar forces may have shaped synonymous codon allocation in the two sets of human sequences, and possibly in the entire human ORFeome. Such suggestion would agree with our previous results **[15]** and with an ongoing analysis of over 3,000 sequences in human and a nematode wherein intercodon dinucleotide tendencies in human coding sequences were also qualitatively very similar.

Codon replacement in sequences

Taking advantage that we had matching coding sequences we explored if it was possible to replace codons in the human sequences by those in *E. coli* sequences. Such codon replacement under equal peptide sequence could reveal to which extent codon tendencies, e.g. intercodon dinucleotides, were dependent on codon usage. Codon substitution was carried out by exchanging pairs of synonymous codons in a mixture of human and *E. coli* sequences as explained under Methodology. The shuffling procedure is schematized **(Figure 5).**

After replacement we found that the codon usage in the "ecolized" human sequences partially differed from those in *E. coli*, for instance, the AGA and AGG codons, which we showed **(Figure 2)** were more frequent in human than in *E. coli*, still had a higher frequency in the changed sequences **(Figure 6).** This showed that codon replacement as performed was unable to completely incorporate the *E. coli* codon usage into human sequences.

With this impediment we did not proceed to other analysis of the codon-converted sequences. However, to test if the obstacle to full codon switch was in some way dependent on the overall codon availability in the donor *E. coli* sequences, we shuffled synonymous codons individually; in this case nearly perfect codon change was achieved **(Figure 6)).** So, codon pairs seem to be a fundamental property of coding sequences in different species. This would agree with the proposal of general and species specific codon-pair context rules **[11].**

Rebuilding a protein-coding sequence in pieces

We also briefly tested an alternative way of codon substitution, namely, the reconstruction of a coding sequence in pieces. We hence used human sequences to rebuild the 238-codon sequence for the green fluorescent protein (GFP) from the jellyfish *Aequorea victoria*. As a way to test the feasibility of the approach we searched for common tetrapeptides, the identified peptides were then used to retrieve the corresponding human coding DNA, duplicated DNA sequences were avoided.

Next, we chose those human coding sequences that matched 59 non-overlapping tetrapeptides that comprised the GFP protein sequence, except for the last two amino acids. We found that multiple human sequences could code for the same tetrapeptide; on the average 54 sequences per tetra-peptide, with a minimum of 6 and a maximum of 113 sequences, except for the tetrapeptide DNHY for which there were only 3 coding sequences.

We then applied the online program WebLogo **[21]** to define a consensus for each tetrapeptide-coding sequence, even though this was not optimal in terms of the imbalance in sequence numbers, we opted for this alternative to avoid an arbitrary selection of sequences. The consensus sequences for each tetrapeptide were then manually ordered to re-create the entire GFP coding DNA. A graphical summary of the procedure is presented in **(Figure 1)** and the codon-converted sequence is given elsewhere **Table 3 (Available with authors).**

Inevitably, codon numbers are small for a 238-amino acid protein, and this complicated analysis of the converted sequence; however, to get a rough idea of the efficiency of the procedure we computed codon usage **(Figure 7)** and intercodon dinucleotide frequencies **(Figure 8)** in the "humanized" GFP. The obtained values approximated more those in the human ORFeome than in GFP, which suggested satisfactory conversion of the GFP coding sequence. Conceivably, analogous automatized reconstruction of coding sequences could swiftly produce codon-converted sequences which might be directly tested for heterologous protein expression.

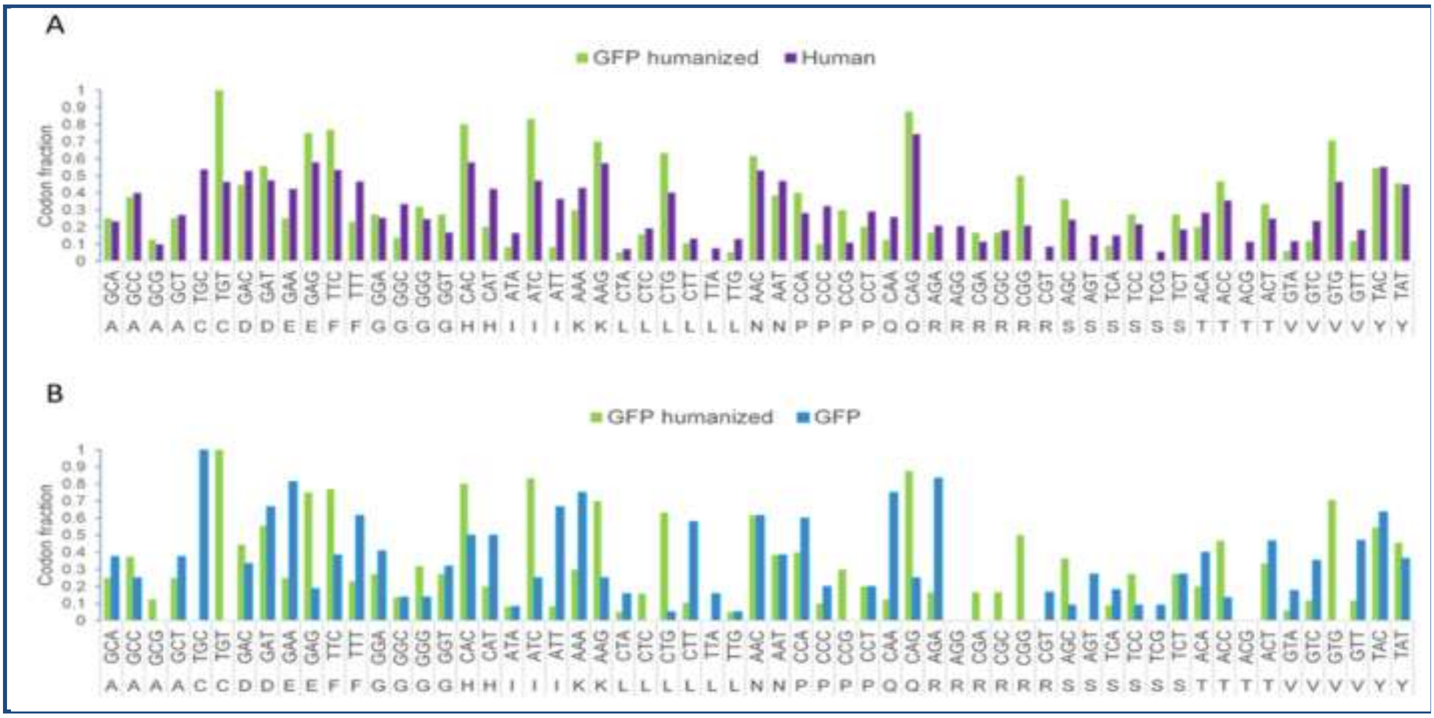

Figure 7: Codon usage in GFP *Aequorea victoria* coding DNA after reconstruction with human coding sequence. Codon usage (Y axis) is expressed as fraction of the unit. In the X-axis codons and corresponding encoded amino acid (one letter code) are shown. Positions where bars are missing indicate absence of that codon either in reconstructed GFP and / or in the original GFP. Codons for Met (ATG), TRP (TGG) and stop codon are omitted. To ease visual appreciation comparisons between converted GFP (GFP humanized) and human and GFP codon usage, we show in panel A the comparison between GFP humanized and human while in panel B we show the comparison between GFP humanized and GFP.

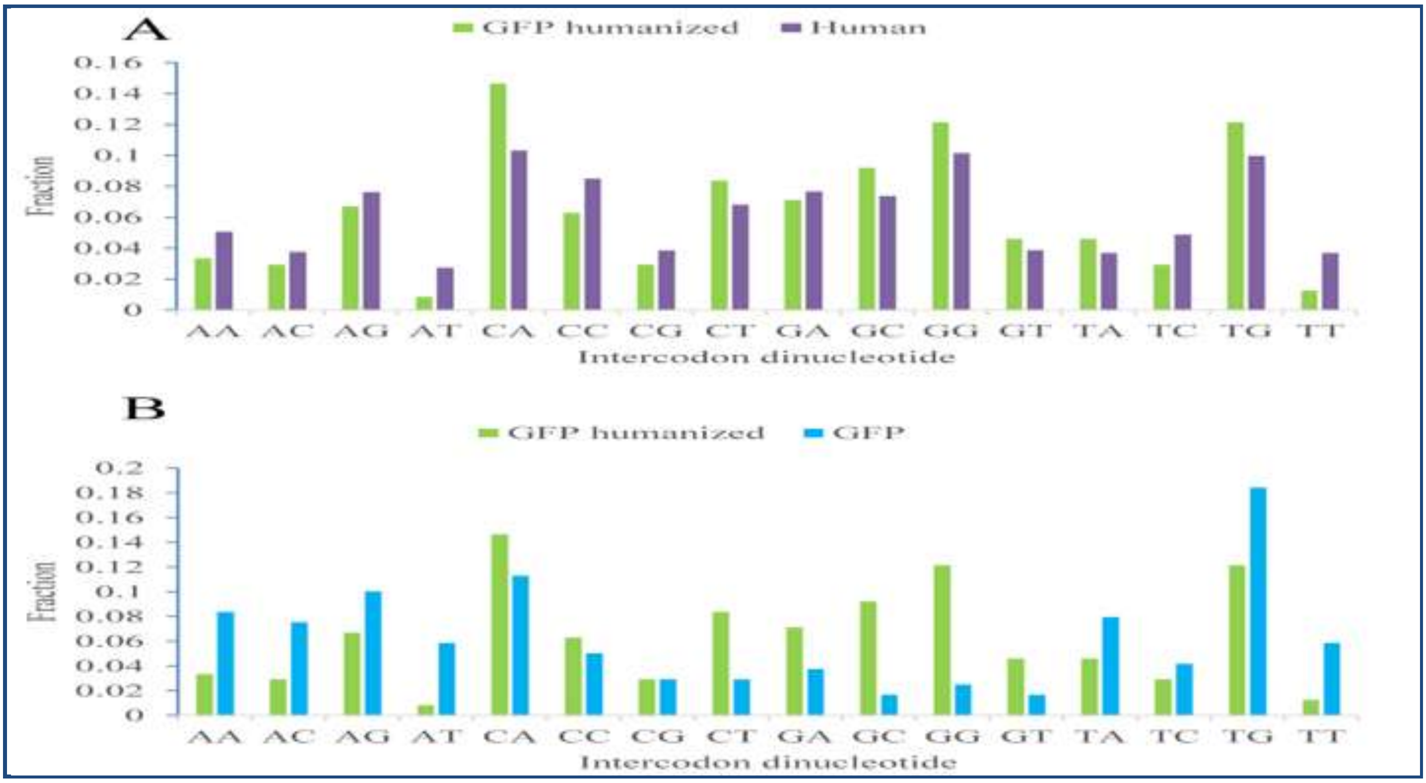

Figure 8: Intercodon dinucleotide frequencies in GFP *Aequorea victoria* coding DNA after reconstruction with human coding sequences. In the x-axis intercodon dinucleotide sequences are shown. In the y-axis intercodon dinucleotide frequency is given as fraction. To ease visual appreciation of comparisons between converted GFP (GFP humanized) and human and GFP, we show in panel **(A)** the comparison between GFP humanized and human and in panel **(B)** the comparison between GFP humanized and GFP.

Conclusions:

The primary goal of this work was to examine the likelihood of using sequences encoding identical peptides to characterize coding DNA from different organisms; results provided proof-of-principle of the proposed approach. We also explored a potential practical utility of the proposed method. Accordingly, replacement of human by *E. coli* codons in sequences coding for the same peptides via the exchange of pairs of synonymous codons was attempted, however, the experiment revealed that exchange of codon pairs precluded full codon replacement. Such results indicate that codon pairs are a primordial genomic feature. Finally, we showed that it is possible to rebuild a protein coding sequence with sequences from a different organism, such procedure could find applicability in heterologous gene expression.

References:

[1] Plotkin J B & Kudla G, *Nat Rev Genet.* 2011 **12**: 32 [PMID: 21102527]
[2] Ikemura T, *Mol Biol Evol.* 1985 **2**: 13 [PMID: 3916708]
[3] Sharp P M *et al. Nucleic Acids Res.* 1988 **16**: 8207 [PMID: 3138659]
[4] Hershberg R & Petrov DA, *PLoS Genet.* 2009 **5**: e1000556 [PMID: 19593368]
[5] Irwin B *et al. J Biol Chem.* 1995 **270**: 22801 [PMID: 7559409]
[6] Qian W, *et al. PLoS Genet.* 2012 **8**: e1002603 [PMID: 22479199]
[7] Li G W *et al. Nature.* 2012 **484**: 538 [PMID: 22456704]
[8] Cannarozzi G *et al. Cell.* 2010 **141**: 355 [PMID: 20403329]
[9] Fedorov A *et al. Nucleic Acids Res.* 2002 **30**: 1192 [PMID: 11861911]
[10] Gutman G A & Hatfield G W, *Proc Natl Acad Sci USA.* 1989 **86**: 3699 [PMID: 2657727]
[11] Moura G *et al. Genome Biol.* 2005 **6**: R28 [PMID: 15774029]
[12] De Amicis F & Marchetti S, *Nucleic Acids Res.* 2000 **28**: 3339 [PMID: 10954603]
[13] Tekaia F *et al. Gene.* 2002 **297**: 51 [PMID: 12384285]
[14] Kahali B *et al. Biochem Biophys Res Commun.* 2007 **354**: 693 [PMID: 17258174]
[15] Sánchez J, *Bioinformation.* 2011 **6**: 327 [PMID: 21814388]
[16] Nussinov R, *J Mol Biol.* 1981 **149**: 125 [PMID: 6273582]
[17] Nussinov R, *J Biol Chem.* 1981 **256**: 8458 [PMID: 6943145]
[18] Pfeifer G P, *Curr Top Microbiol Immunol.* 2006 **301**: 259 [PMID: 16570852]
[19] Chen GT & Inouye M, *Genes Dev.* 1994 **8**: 2641 [PMID: 7958922]
[20] Trotta E, *PLoS One.* 2011 **6**: e21590 [PMID: 21738721]
[21] Crooks G E *et al. Genome Res.* 2004 **14:** 1188 [PMID: 15173120]

Edited by P Kangueane
Citation: **Sánchez *et al.*** Bioinformation 9(10): 511-517 (2013)