

# Recognition Sequences in the Restriction Endonucleases.

Jan C. Biro<sup>1,2</sup>, Josephine M.K. Biro<sup>2</sup>

<sup>1</sup>Karolinska Institute, Stockholm, Sweden.

<sup>2</sup>Homulus Informatics, 88 Howard, # 1205, San Francisco, 94 105 CA., USA.

[jan.biro@sbcglobal.net](mailto:jan.biro@sbcglobal.net)

## Abstract

The nature of specific protein-nucleic acid interaction between restriction endonucleases (RE) and their recognition sequences (RS) was studied by bioinformatics methods. It was found that the frequency of 5-6 residue long RS-like oligonucleotides is unexpectedly high in the nucleic acid sequence of the corresponding RE ( $p < 0.05$  and  $p < 0.001$  respectively,  $n=7$ ). There is an extensive conservation of these RS-like sequences in RE isoschizomers. A review of the seven available crystallographic studies showed that the amino acids coded by codons that are subsets of recognition sequences were often closely located to the RS itself and they were in many cases directly adjacent to the codon-like triplets in the RS. Ten examples of this codon - amino acid co-localization are presented. The distance between the nitrogen and oxygen atoms of the co-localized molecules is short,  $3.74 \pm 0.46$  Å (mean  $\pm$  S.E.), indicating that an interaction between the nucleic- and amino acids might occur.

**Keywords:** restriction endonuclease, recognition sequence, protein - nucleic acid interaction, codon, amino acid, molecular structure, isoschizomer,

## 1 Introduction

The nature of specific protein – nucleic acid interactions is not well understood. The interaction between transcription factors and promoters has been studied most extensively. However, that system is very complex and there has been, so far, no simple, general conclusion drawn from these studies. The interaction of restriction enzymes (REs) with their recognition sequences (RSs) is also highly specific. Furthermore, the protein-binding site is often short (5-7 nucleotides) and simple (tandem repeat, where the sense and anti-sense strands are identical). The RE-RS system is also extensively studied because of its great biotechnological importance. These circumstances makes the RE family an interesting case study of specific nucleic acid – protein interactions.

Our previous study [1] convinced us that the codon translation table is not random and that there is a common

periodicity in the codon structure and the physico-chemical properties of the amino acids. We interpreted those results in favor of Woese who argued [2] that the genetic code developed in a close connection to the amino acid repertoire and that this close biochemical connection is fundamental to specific protein – nucleic acid interactions. This consideration led us to ask whether the genetic code could be somehow the bridge between a nucleic acid sequence (here the RS) and the amino acid sequence (here the RE) that specifically recognizes it.

## 2 Materials and methods

Restriction enzyme data was collected from REBASE [3], GenBank [4], SwissProt [5] and the Protein DataBank (PDB) [6]. Nucleic acid and protein sequences were aligned and compared to each other using ClustalW [7] and the similarities were visualized using Jalview [8]. In some cases the nucleic acid sequences were overlappingly translated into virtual protein-like sequences [9]. We found that overlappingly translated sequences (OTS) are especially useful for detecting and visualizing short sequence similarities (in contrast to regularly translated proteins, unpublished) because they retain all the information present in the nucleic acid sequences, while the regular, non-overlapping translation loses as much as  $2/3^{\text{rd}}$  of information because of codon redundancy.

This study was limited to those restrictions endonucleases whose recognition sequence was unambiguous and where sequence and structure data of the DNA-enzyme complex was publicly available. One thousand residue long repeating sequences were constructed from the RS-s. These artificial repeats were compared to the RE nucleic acid sequences using ClustalW to find RS-like oligonucleotides. This method found most, but not all, RS-like sequences. In some cases it was necessary to complete this approach with searching using text search tools and counting of 4-8 nucleic acid long, RS-like oligonucleotides.

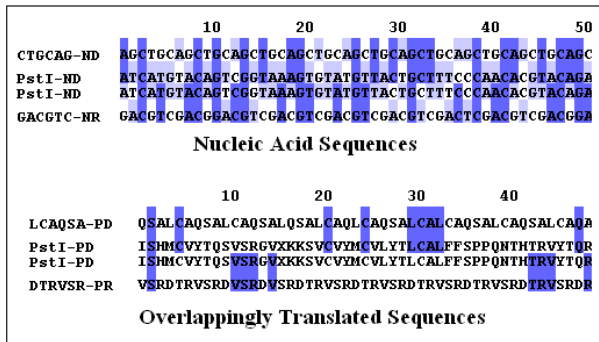
The crystallographic structures were visualized and analyzed using Swiss-PdbViewer [10]. The student's  $t$ -test was used for statistical evaluation of the results [11].

## 3 Results

The restriction enzyme PstI has 11 cloned and sequenced isoschizomers (restriction enzymes that recognize the same DNA sequence; the cut sites may or may not be identical.). They all specifically recognize the sequence CTGCAG in direct (D) reading; this is identical to its

reversed and complemented (RC) sequence. The reverse (R) and complementary (C) readings are GACGTC. These short sequences were repeated 167 times to form two about 1000 residue long repeats of this recognition site, called CTGCAG-ND-1000 and GACGTC-NR-1000. When RS-repeats were aligned to the RE, using the ClustalW program, many short RS-like sequences were found in the RE-coding DNA (**Figure 1**). However the nucleic acid alignment turned out to be very “noisy” because of many identical single nucleotides. Therefore both the RE and RS nucleic acids were overlappingly translated and the OTS sequences were aligned using ClustalW. This approach effectively filtered the single nucleotide similarities. Neither the nucleic acid nor the OTS alignment found all RS-like residues and the two approaches gave slightly different results (different nucleotides in the last wobble positions are often interpreted to code for the same amino acid).

### ClustalW Alignment of the PstI and Recognition Sequences



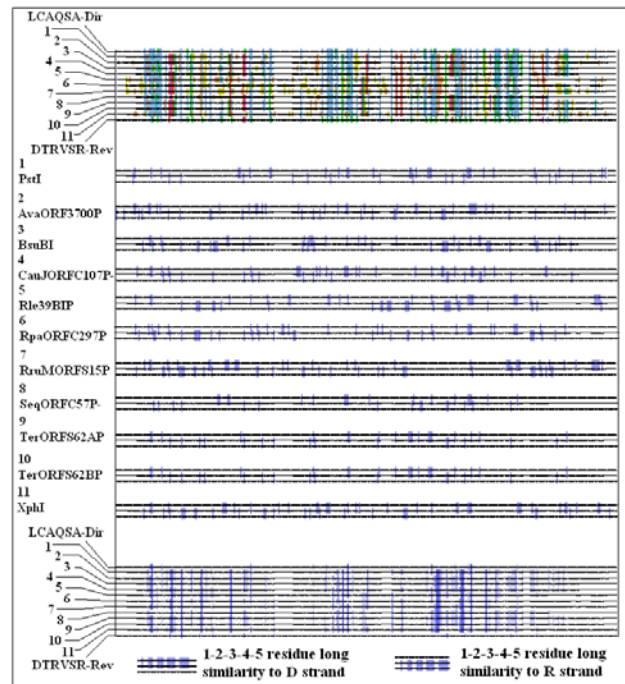
**Figure 1: ClustalW Alignment of PstI Coding Sequences and the PstI Recognition Sequence.** The PstI nucleic acid sequence was compared to PstI RS repeats (direct, D and reverse, R readings) as nucleic acids (N) and as Overlappingly Translated Sequences (P). The result of the alignments is visualized by Jalview where the shaded areas emphasize sequence identity. The consensus calculations are also shown above and below the sequences (Quality). Only residues 801-850 of the enzyme are shown.

Multiple sequence alignment (MSA) of the 11 PstI isoschizomers showed that these enzymes are similar to each other, as was expected. A simultaneous MSA involving the RS-like repeats showed that a substantial number of the similarities between enzymes are caused by common short sequence similarities to their common RS (**Figure 2**). It was found that even the reverse RS-like sequences are represented in the REs. A large number of RS-like sequences are present in the majority of the RE sequences at the same position. This conservation indicates that the involved residues are significant even if they are only 1-2 OTS letters (3-4 nucleic acids) long.

It was necessary to study the known three-dimensional structures of the RS-RE complexes to understand the biological meaning of the presence of RS-like sequences in the REs and its possible effects on the specific DNA-protein interaction. Crystallographic data for seven different REs was available in July 2003 (**Table I**).

These enzymes (except two) are not isoschizomers. The nucleic acid sequence of each was aligned to its own RS

### Recognition Sequence -like Sites in the Restriction Enzymes



**Figure 2.: Recognition Sequence - like Sites in the Restriction Enzymes:** Multiple Sequence Alignment (MSA) of 11 REs (PstI isoschizomers) and their common RS. The coding sequences of the enzymes and the direct (Dir, D) and reverse (Rev, R) readings of the poly-RSs were Overlappingly Translated before the MSAs were performed. The first and last alignments include all 13 sequences (colored by ClustalW colors, first and by conservation, last) while the middle 11 alignments indicate individual comparisons of the enzymes to their common RS (colored by conservation). The section of the alignments seen corresponds to residues 1601-1904 of PstI .

**Table I**  
Restriction Enzymes with Known Crystallographic Structure

NAME	Recognition Sequence (RS)	Codon Potential	Crystal Name	AC# Gene-Bank	N.A.-residue #	AC# Swiss-Prot	A.A.-residue #	RS-like oligonucleotides (- # of copy)
BanHI	G'GATCC	GDISPRLA	1BHM	X55285	642	P23940	213	GGCCTA-1, GCCTA-1, TAGG-1, GGAT-2, AGGC-1, GATC-1.
EglII	A'GATCT	RNLSLX	1DFM	U49842	672	-	223	AGATC-1, TAGAT-1, AGAT-5, GATC-2, CTAG-2.
EcoRI	G'AAATTC	ENIPSR	1ERI	J01675	909	P06642	302	CGAATT-1, AGCTTA-1, TTCGA-2, AAGCT-3, AATTC-2, AATT-3, TTAA-6, TAAG-3, CTTA-3, TCGA-3, GAAT-4, AAGC-3, TTCG-1.
EcoRV	GAT'ATC	NIPSR	LAZO	X00530	738	P04390	245	ATATCG-1, ATATC-1, GATAT-3, TATAG-1, ATAT-11, TATC-2, GATA-2, TATA-4, CTAT-2.
NaeI	GCC'GGC	APRG	1IAW	U09881	954	-	317	GGCGCCGG-1, GCCCGGG-1, CGGCGC-2, GCCCGG-1, GCGCG-2, GCGCG-1, CGGCG-2, GCGCG-2, GGCGG-3, CGGC-1, CGGC-3, CCGG-5, GCGG-3, CCGG-4, GCGG-3, GGCC-1, CGCG-4.
NgoMIV	G'CCGGC	APRG	1FTU	M86915	861	P31032	286	CCGGC-1, GCGG-4, GCGG-1, GCGG-3, CGCC-2, CCGG-1, GCGG-2, CGGC-2.
PvuII	CAG'CTG	QSALCA	1PVI	AF305615	474	P23657	157	AGCT-2, GCTG-1.

AC#: accession number, N.A.: nucleic acid, A.A.: amino acid, ' cut site

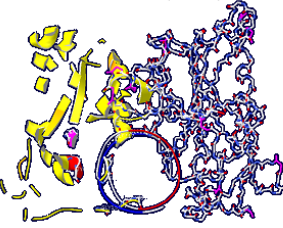
– repeat. (OTS were not used in this study.) The results of the ClustalW alignments were manually checked and completed. The number and position in the RE sequence of each RS-like oligonucleotide longer than 3 residues was counted and recorded. The amino acids that corresponded most closely to these oligonucleotides (using the regular, 3-letter, non-overlapping codon table) were localized in the protein sequences and 3D structures of the enzymes. The RS-like sequences found using this

method are summarized in **Table I**. The locations of the corresponding amino acids are illustrated in **Figure 3**.

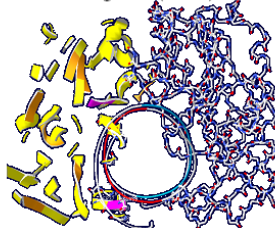
The statistical evaluation of the results was based on the calculation of the number of strings which are expected (**E**) to be found in a **L** residues long sequence only by chance and compare this number with the number of the same strings that are really found (**F**). The formula  $E=L/4^n$  was used, where **n** is the number of residues in the string. The result of this statistical evaluation is shown in **Figure 4**. The first 2 bars of the figure (marked by \*) require additional explanation ; this also gives an example of our calculations. We have found one 7-residue and one 8-residue long RS-like sequence in *NaeI*, which is 954 residues long. The expected values are  $954/4^7=0.058$  and  $954/4^8=0.014$  respectively, whereas  $F=1$  in each case. Thus the  $F/E$  ratios are 17.2 and 71.4 respectively, indicating that these findings are significant although it was not possible to use the student's t-test on these single values.

### The Location of RS-like Sequences in the 3D Structure of REs

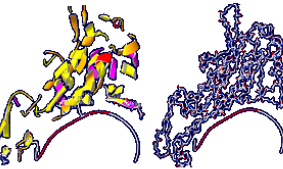
a. Structure of BamHI-RS (LBHM)



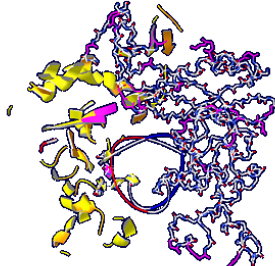
b. Structure of BglIII-RS (IDFM)



c. Structure of EcoRI-RS (IERD)



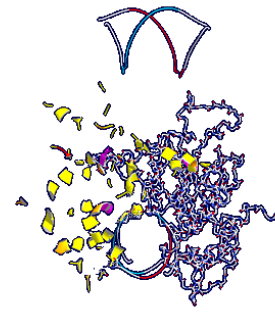
d. Structure of EcoRV-RS (IAZO)



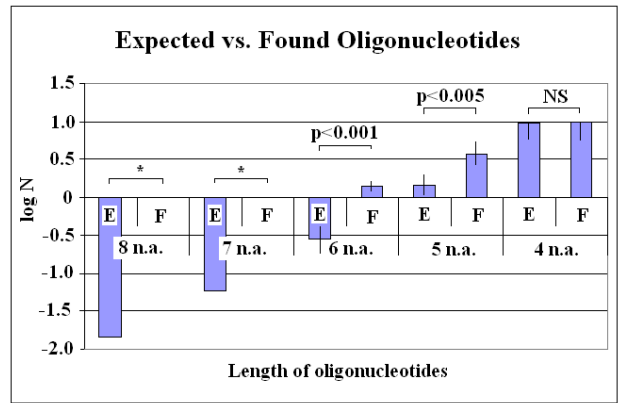
e. Structure of NaeI-RS (IIAV)



f. Structure of NgoMIV-RS (LFIU)



**Figure 3: The Location of RS-like Sequences in the 3D Structure of REs.** The figure shows the characteristic structure of REs: two identical subunits around dsDNA (the RS). One subunit is left intact (continuous solid 3D rendering of the backbone). The other subunit is used to indicate the amino acids that are coded by nucleic acids corresponding to the RS-like sequences in their coding DNA (interrupted ribbon structure). The color code of the ribbon backbone indicates the length of the RS-like strings: yellow = 3, orange = 4, pink = 5, red  $\geq 6$  (strings  $< 3$  residues long are not indicated). The solid spirals indicate the dsDNA; the red and blue lines are the RSs while the white parts are not RSs. (The EcoRI structure is an exception, there is only a single DNA strand and only one enzyme subunit).

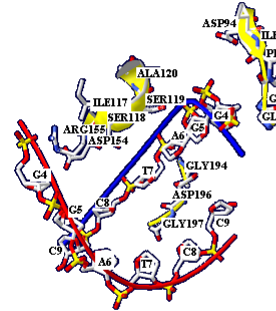


**Figure 4: Expected (E) vs. Found (F) RS-like Oligonucleotides in the REs.** Expected and observed numbers (N) of RS-like nucleotides from 4 to 8 residues long are shown. Statistically significant E - F differences are indicated. NS: not significant, \*: single value. For details see the *Results*

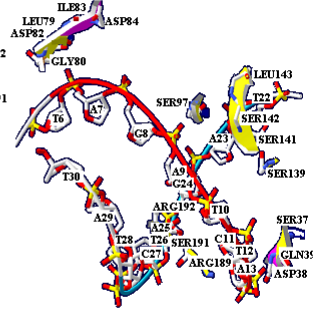
The distribution of amino acids related to 3-4 residue long RS-like sequences in the REs seems to be rather even, however there is a tendency for the amino acids that are related to longer (5-8 long) RS-like oligonucleotides to be located close to the DNA. A substantial number of the amino acids that are located in grooves of the RS-DNA are coded by RS-like codons (**Figure 5**).

### Amino Acids Coded by RS-like Codons and Co-located with RS.

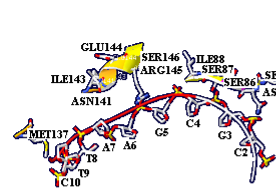
BamHI (LBHM) - RS with 14 a.a.



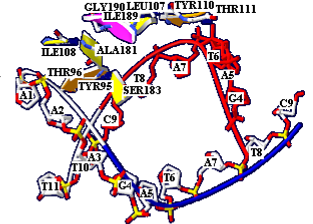
BglIII (IDFM) - RS with 16 a.a.



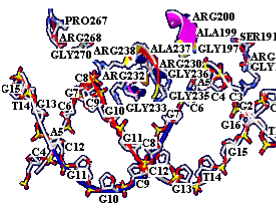
EcoRI (IERD) - RS with 11 a.a.



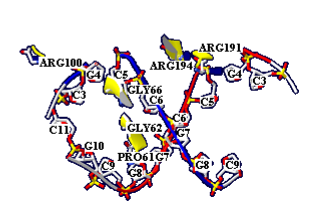
EcoRV (IAZO) - RS with 10 a.a.



NaeI (IIAV) - RS with 16 a.a.



NgoMIV (LFIU) - RS with 6 a.a.



**Figure 5: Amino Acids in REs Coded by RS-like Codons and Co-located with RSs.** The color code of the ribbon backbone indicates the length of the RS-like strings: yellow = 3, orange = 4, pink = 5, red  $\geq 6$ . The solid spirals indicate the dsDNA (with the phospho-deoxyribose backbone); the red and blue lines are the RSs while the white lines are not RSs. (EcoRI is an exception, there is only a single DNA strand). a.a.: amino acid

It was possible to find many examples where an amino acid was co-located with its codon-like triplet in the RS (Figure 6).

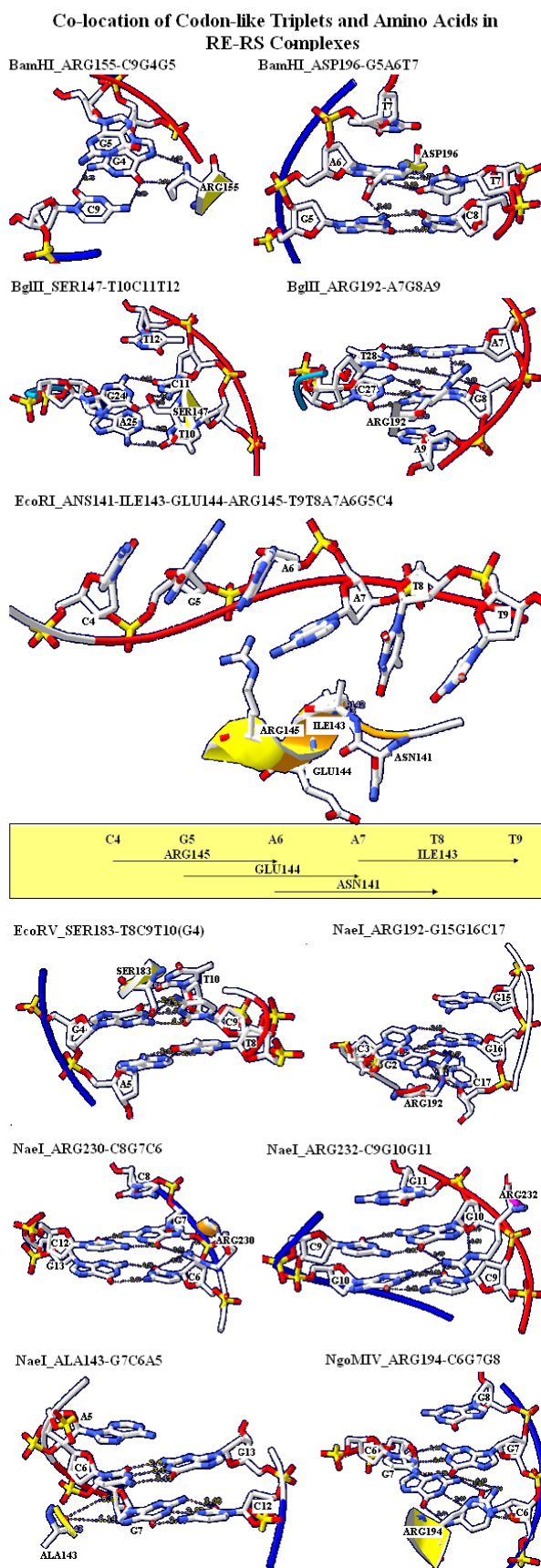


Figure 6.: Co-location of Codon-like Triplets and Amino Acids in RE-RS Complexes. Examples are taken from Figure 5.

In ten cases the nitrogen (N) or oxygen (O) atoms in the amino acid residue were within direct or indirect hydrogen bonding distance to an O or N atom in the first or second nucleotide residue of its codon-like triplet in the RS. These amino acids, the codon-like triplets and the distance between the closest atoms are listed in the Table II. The average distance between the atoms indicated is  $3.74 \pm 0.36$  Å (mean  $\pm$  S.E.,  $n=10$ ). These distances are short enough to indicate interactions (probably through H-bridges) between the molecules. We have found many examples where an amino acid was co-located with its codon-like triplet in the RS but without interaction with the nucleic acid bases. In these cases the amino acid residues were aligned along the phosho-deoxyribosyl backbone of the DNA, close to the O atoms in the phosphate groups. A rather interesting example for this type of molecular alignment was found in EcoRI (part of Figure 6).

In this example all the four theoretically possible overlappingly translated amino acids of the sequence CGAATT were co-located with the RS (GAATTC).

Only examples of supposed interactions between the amino acid residue and nucleic acid bases are listed here and the numerous examples of interactions involving the nucleic acid phosphate-deoxyribose backbone are not shown.

Table II

Codon - Amino Acid Co-location: the Shortest Atomic Distances

Restriction Enzyme	Amino Acid	Codon	Closest nucleotide	Closest atoms	Shortest Distance (Å)
BamHI	ARG155	C9G4G5	G4	N - O	2.59
	ASP196	G5A6T7	G5	O - N	3.08
BglII	SER145	T10C11T12	C11	O - N	4.44
	ARG192	A7G8A9	G8	N - O	3.85
EcoRV	SER183	T8C9T10	C9	O - N	4.89
NaeI	ARG192	G15G16C17	C17	N - O	2.76
	ARG230	C8G7C6	G7	N - O	2.61
	ARG232	C9G10G11	G10	N - O	2.78
NgoMIV	ALA143	G7C6A5	C6	O - N	5.87
	ARG194	C6G7G8	G7	N - O	4.49
Mean $\pm$ S.E.					3.74 $\pm$ 0.36

## 4 Discussion

Specific DNA-protein interactions are very important in the regulatory network of the genome. The exact rules of these interactions are not well understood. The known forms of DNA (the Double Helix) are closed, inverted structures where the molecular information is not directly exposed on the surface [12]. However the major groove is rich in chemical information. The edges of each base pair are exposed in the major and minor grooves, creating a pattern of hydrogen bond donors (**D**) and acceptors (**A**) and of van der Waals surfaces (methyl group, **M**; nonpolar hydrogen, **H**) that identifies the base pair [13]. There is a unique and logical link between 1., the nucleotide sequence of the DNA that specifically interacts with a protein; 2., the pattern of **D**, **A**, **M**, **H** properties in the grooves of that DNA sequence; 3., the physicochemical properties of the protein that interacts with it and 4., the DNA coding the amino acids of that protein. The question is whether this unique and logical link is the *genetic code* itself.

This question was already formulated in the 1960s and there are basically two distinct opinions. Francis Crick could not see any logical connection between the structure of the genetic code and the physicochemical properties of the amino acids and he regarded it just a “frozen accident” [14]. On the other side, Woese [2] propagated the theory of the coevolution of proteins and nucleic acids and argued for a specific stereochemical connection between the amino acids and their codons. We succeeded in constructing a “Common Periodic Table of Codons and Amino Acids” [1] (Biro et al, 2003) and so became fellows of Woese.

The REs are known to interact very specifically with their RSs. We tried to find RS-like oligonucleotides in the coding sequences of the REs. The RSs are usually simple, short sequences, and it is not possible to find 3-6 residue long sequences by using conventional sequence similarity searching methods such as BLAST or FASTA. However, an unconventional method, the multiple sequence alignment of overlappingly translated sequences, seems to be useful for finding and visualizing short sequence similarities. The method is rapid and informative. A disadvantage is that there are no methods developed for exact statistical evaluation of the results. We were able to confirm that PstI isoschizomers are rather similar to each other and contain conserved sequences (as expected). However, we also made the new observation, that many of these sequence conservations are short, conserved, RS-like sequences. Even if some of the shortest (3-4 nucleic acid long) RS-like sequences could easily have been found by chance, the conservation indicates biological significance.

This indication was further strengthened by our second study of seven REs with known 3D structures. A statistically significant overrepresentation of 5-8 residue long RS-like sequences were found in the coding sequences of these enzymes. Conservation and overrepresentation do not automatically confirm a biological role, however it is a strong argument for one [15]. Codons for alanine, glycine, valine and aspartate have relatively high frequency in the acceptor stems of their respective tRNAs [16]. The tRNAs with complementary anticodons also had some kind of complementarity with their acceptor stems [17]. Such relationships could support the hypothesis that one or more anticodon nucleotides were historically related to an acceptor stem nucleotide needed for aminoacylation, i.e. they are signs of codon – amino acid co-evolution.

Even more convincing evidence is, of course, the visualization of a stereochemical relationship between a particular codon and its amino acid. The increasing amount of freely available crystallographic data, including structures of DNA-protein complexes, might give us this type of evidence and we show here an example of it. We were able to find ten examples where a nucleic acid was co-located with its own codon in such a way that it might indicate a stereospecific interaction. In each example atoms with opposite partial charge (N, O) were involved, they were close enough to each other to interact, and the first or second bases but never the third (wobble) base of the codon were co-located with the

amino acid side-chain. However, the number of examples is few, the majority of the involved amino acids are charged and there is a particular over-representation of Arginine. We are aware of some early model studies [18, 19] indicating stereochemical relationship between coding triplets and amino acids, as well as the error in that model building [20]. We don't want to repeat that mistake while searching for a fallen apple close to its tree.

## 5 Acknowledgement

The author is grateful to Dr Clare Sansom (Birkbeck College, London) for her helpful comments and suggestions regarding the preparation of the manuscript.

## 6 References

- [1] J.C. Biro, B. Benyo, C. Sansom, A. Slavecz, G. Fordos, T. Micsik, Z. Benyo, A common periodic table of codons and amino acids. *Biochem. Biophys. Res. Com.* 306 (2003) 408-415.
- [2] C.R. Woese, in: *The Genetic Code: The Molecular Basis for Gene Expression*, Harper & Row, New York, 1967, pp. 156-160, Chapters 6-7.
- [3] R.J. Roberts, T. Vincze, J. Posfai, D. Macelis. REBASE - restriction enzymes and methylases. *Nucleic Acids Research* 31 (2003) 418-420. - <http://rebase.neb.com>
- [4] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler. GenBank. *Nucleic Acids Res.* 31 (2003) 23-7. <http://www.ncbi.nlm.nih.gov/>
- [5] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, L. Phan, S. Pilbout, M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31 (2003) 365-370. - <http://us.expasy.org/sprot/>
- [6] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28 (2000) 235-242. - <http://www.rcsb.org/pdb/index.html>
- [7] J.D. Thompson, D.G. Higgins, T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22 (1994) 4673-80 - <http://www.ebi.ac.uk/clustalw/#>
- [8] M. Clamp. Jalview. 1999 - <http://www.ebi.ac.uk/~michele/jalview/>
- [9] J.C. Biro Overlapping translation of nucleic acids for bioinformatics applications. *Med. Hypotheses* 60 (2003a) 654-659.
- [10] N. Guex, M.C. Peitsch. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18 (1997) 2714-2723. - <http://www.expasy.org/spdbv/>

- [11] Student's t-test -  
[http://www.physics.csbsju.edu/stats/t-test\\_NROW\\_form.html](http://www.physics.csbsju.edu/stats/t-test_NROW_form.html)
- [12] J.C. Biro. Speculation about alternative DNA structures. *Med. Hypotheses* 61 (2003/c) 86-97.
- [13] J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine, R. Losick. in *Molecular biology of the gene* (5<sup>th</sup> edition): The structure of DNA and RNA. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2004, pp 1-33 chapter 6 (preprint in 2003).
- [14] F.H.C. Crick. The origin of the genetic code. *J. Mol Biol.* 38 (1968) 367-379.
- [15] P. Schimmel. Origin of genetic code: A needle in the haystack of tRNA sequences. *Proc. Natl. Acad. Sci. USA* 93 (1996) 4521-4522.
- [16] W. Moller, G.M. Janssen. Statistical evidence for remnants of the primordial code in the acceptor stem of prokaryotic transfer RNA. *J. Mol. Evol* 41 (1992) 471-477.
- [17] S. Rodin, A. Rodin, S. Ohno. The presence of codon-anticodon pairs in the acceptor stem of tRNAs. *Proc. Natl. Acad. Sci. USA* 93(1996) 4537-4542.
- [18] S. R. Pelc, M.G.E. Welton Stereochemical relationship between coding triplets and amino-acids. *Nature* 209 (1966) 868-870.
- [19] M.G. E. Welton. S. R. Pelc. Specificity of the stereochemical relationship between ribonucleic acid-triplets and amino-acids. *Nature* 209 (1966) 870-872.
- [20] F. H. C. Crick. An Error in Model Building. *Nature* 213 (1967) 798.