Research article

Frequent occurrence of recognition Site-like sequences in the restriction endonucleases

Jan C Biro*[1,2] and Josephine MK Biro[2]

Address: [1]Karolinska Institute, Stockholm, Sweden and [2]Homulus Informatics, 88 Howard, # 1205, San Francisco, 94 105 CA, USA

Email: Jan C Biro* - jan.biro@sbcglobal.net; Josephine MK Biro - jan.biro@sbcglobal.net

* Corresponding author

Abstract

Background: There are two different theories about the development of the genetic code. Woese suggested that it was developed in connection with the amino acid repertoire, while Crick argued that any connection between codons and amino acids is only the result of an "accident". This question is fundamental to understand the nature of specific protein-nucleic acid interactions.

Results: 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.

Fifty-five examples of this codon-amino acid co-localization are found and analyzed, which represents 41.5% of total 132 amino acids which are localized within 8 Å distance to the C1' atoms in the DNA. The average distance between the closest atoms in the codons and amino acids is 5.5 +/- 0.2 Å (mean +/- S.E.M, $n = 55$), while the distance between the nitrogen and oxygen atoms of the co-localized molecules is significantly shorter, (3.4 +/- 0.2 Å, $p < 0.001$, $n = 15$), when positively charged amino acids are involved. This is indicating that an interaction between the nucleic- and amino acids might occur.

Conclusion: We interpret these results in favor of Woese and suggest that the genetic code is "rational" and there is a stereospecific relationship between the codes and the amino acids.

Background

The nature of specific protein-nucleic acid interactions is not well understood. The interaction between transcription factors and promoters has been studied most extensively [1]. 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 [2] 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 [3] 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.

Methods

Restriction enzyme data was collected from REBASE [4], GenBank [5], SwissProt [6] and the Protein DataBank (PDB) [7]. Nucleic acid and protein sequences were aligned and compared to each other using ClustalW [8] and the similarities were visualized using Jalview [9]. In some cases the nucleic acid sequences were overlappingly translated into virtual protein-like sequences [10]. 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 looses as much as 2/3rd 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 [11]. The paired student's t-test was used for statistical evaluation of the results [12].

Results

The restriction enzyme PstI has 12 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 we preferred to use protein sequences instead of nucleic acid sequences. The RS nucleic acids were overlappingly translated and the protein-like OTS sequences were aligned with the usual (non-OTS) protein sequences of the REs using ClustalW. This approach effectively filtered the single nucleotide similarities. The OTS sequence of the RS-repeats provided a frame independent, protein-like representation of the recognition sites and the similarity to the RE sequence indicated the presence of RS-like sequences in the REs. 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).

Multiple sequence alignment (MSA) of the 12 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.

The statistical evaluation of this MSA result confirmed the significance of the common conserved residues in the RSs end REs. (Figure 3). The number of conserved residues in the RE/RS-like alignment was 9.5 +/- 0.6 (RS-D, mean +/- S.E.M., n = 14) and 5.9 +/- 0.6 (RS-R). This number became significantly lower when the RS was randomized in the non-RS-like/RE alignment (2.1 +/- 0.1) or when the REs were replaced by non-RE proteins in the RS-like/non-RE alignment (2.0 +/- 0.1).

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 PDB version (Table 1).
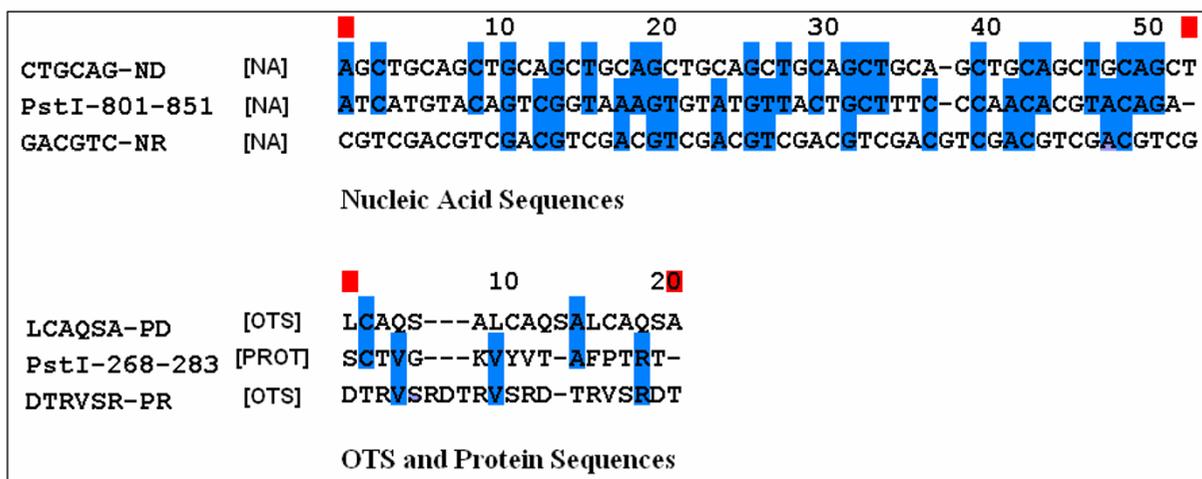
Figure 1
ClustalW alignment of PstI sequences and the PstI recognition sequence. The PstI sequence was compared to PstI RS repeats (direct, D and reverse, R readings) by two different ways: as nucleic acids [NA, N] or as overlappingly translated [OTS, P] and regularly translated [PROT] sequences. The result of the alignments is visualized by Jalview where the shaded areas emphasize sequence identity. Only a short fragment of the alignments is shown.

These enzymes (except two) are not isoschizomers. The nucleic acid sequence of each was aligned to its own RS-repeat. (OTS was 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 oligonucleotides 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 1. The locations of the corresponding amino acids are illustrated in Figure 4.

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 5.

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 = l 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 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 complex are coded by RS-like codons (Figure 6).

It was possible to find many examples where an amino acid was co-located with its codon-like triplet in the RS. An amino acid and a codon was regarded to be co-located if any atom in the amino acid was within 8 Å distance from at least one Cl' atom in the triplet (The Cl' atom is the junction site between the deoxyribose and the nucleic acid base. Compare this value with the diameter of the dsDNA which is about 20 Å). One hundred thirty-two amino acids were located within 8 Å distance from one strand of the RS. Fifty-five of these (41.5 +/- 3.6 %) were co-located with its entire codon (3 letters, ABC) while 27 with codon fragments (AB or BC) and 50 with non-codon triplets. Control experiments, where the co-location with triplets in the randomised RSs was studied showed that
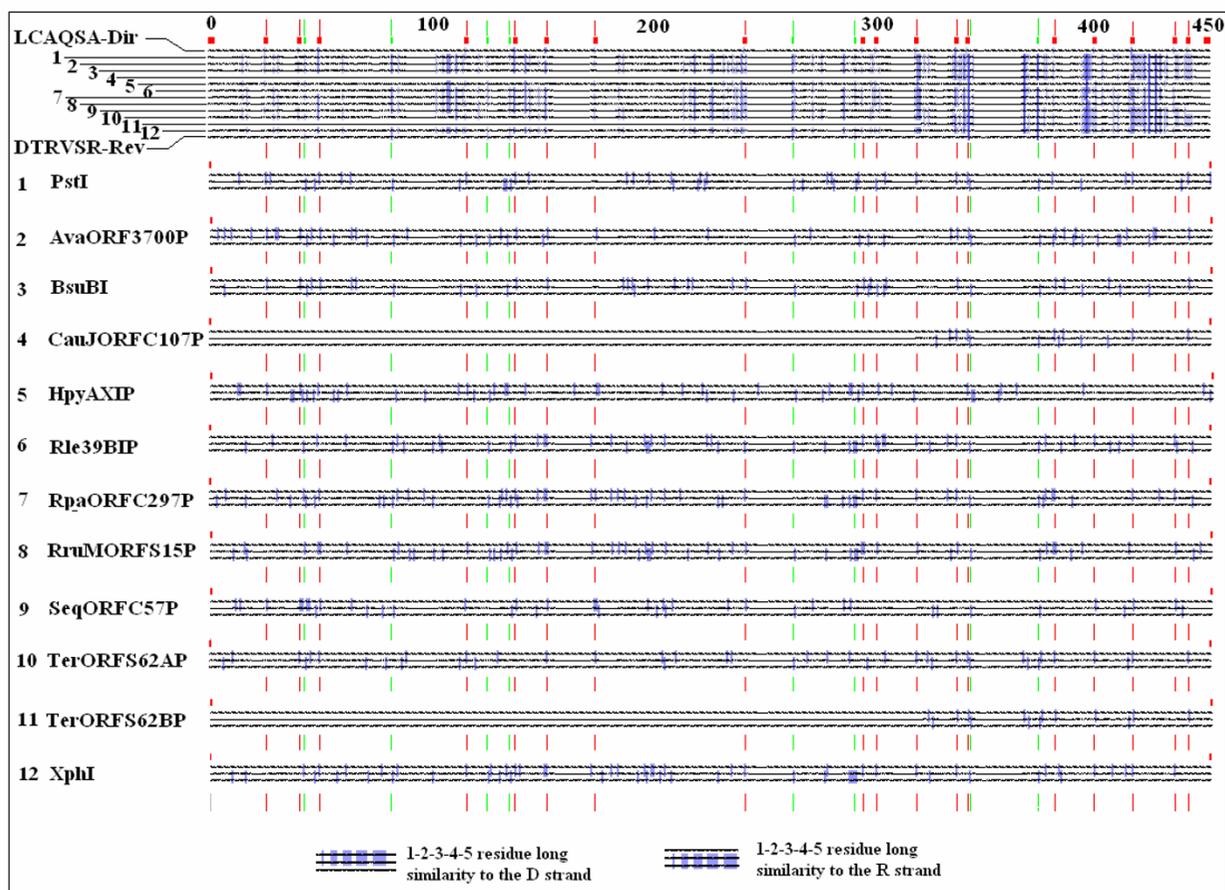
Figure 2
Recognition sequence-like sites in the restriction enzymes: Multiple Sequence Alignment (MSA) of 12 REs (PstI isoschizomers) and their common RS. The protein sequences of the enzymes and the overlappingly translated sequences (OTS) of the RS direct (Dir, D) and reverse (Rev, R) readings were compared. The first alignment includes all 14 sequences (colored by conservation) while the other 12 alignments indicate individual comparisons of the enzymes to their common RS (colored by conservation). The section of the alignments seen corresponds to the entire protein sequence (326 amino acids) of PstI. The vertical lines between the alignments indicate the conserved residues which align to the direct (red) and reverse (green) readings of the RS.

the 55 amino acid co-location to the entire codon triplets were statistically highly significant ($p < 0.001$, n = 7). (Table 2).

The average distance between the closest atoms in the amino acids and codons was 5.5 +/- 0.2 Å (mean +/- S.E.M., n = 55). There was little variation between restrictions enzymes regarding this value and only the arginine rich NaeI showed shorter average distance. The physicochemical properties of the amino acids had significant influence on their distance to their codons. The positively charged amino acids (Arg, Lys, Gln) were closest to the codon (3.4 +/- 0.2 Å, n = 15) while the hydrophobic amino acids were most distantly located (7.3 +/- 0.3 Å, n = 15). (Table 3).

Examples of the different kinds of codon-amino acid co-localizations are shown in Figure 7, 8. It was possible to find examples for 12 of 20 different amino acids. In many 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 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-
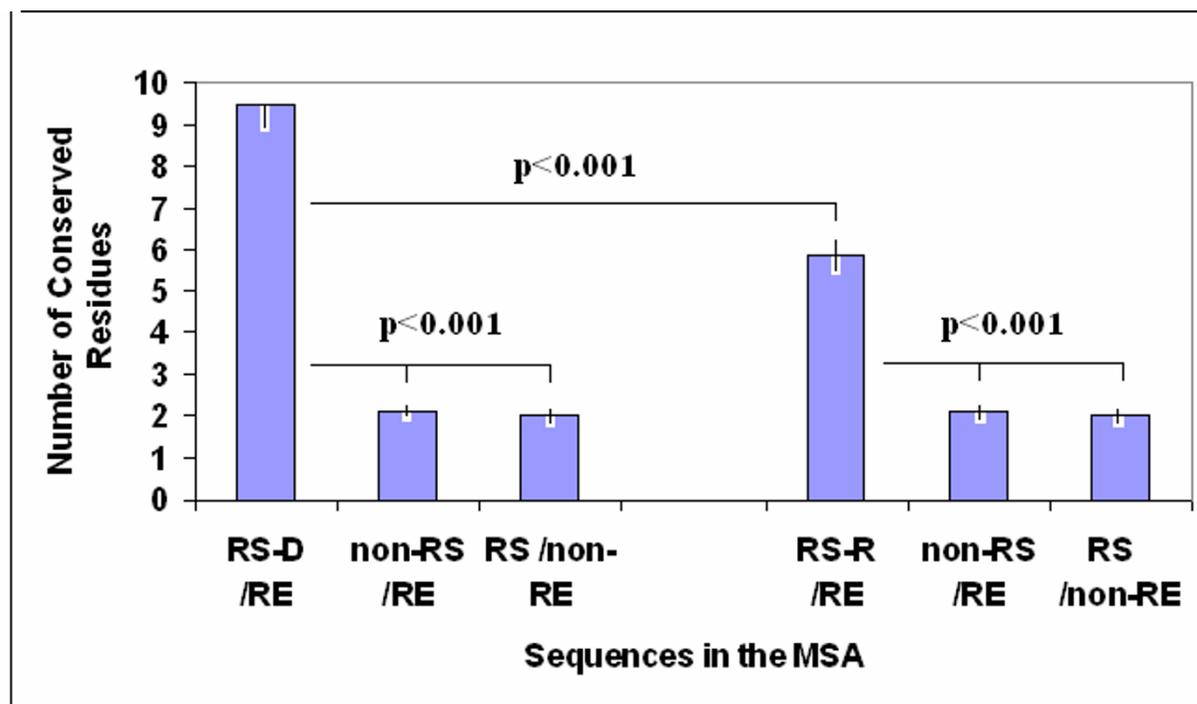
Figure 3
Statistical evaluation of MSA (RS/RE) The number of conserved residues shown in Figure 2 was counted (RS-D/RE and RS-R/RE). In the control experiments similar alignments were constructed but the RS was replaced by shuffled RS (non-RS/RE) or the RE was replaced by shuffled RE (RS/non-RE). Each bar represents the mean +/- S.E.M, n = 12.

Table 1: 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)
BamHI	G†GATCC	GDISPRLA	1BHM	X55285	642	P23940	213	GGCCTA-1, GCCTA-1, TAGG-1, GGAT-2, AGGC-1, GATC-1
BglII	A†GATCT	RNISLX	1DFM	U49842	672	-	223	AGATC-1, TAGAT-1, AGAT-5, GATC-2, CTAG-2
EcoRI	G†AATTC	ENIPSR	1ERI	J01675	909	P00642	302	CGAATT-1, AGCTTA-1, TTCGA-2, AAGCT-3, AATTC-2, AATT-8, TTAA-6, TAAG-3, CTTA-3, TCGA-3, GAAT-4, AAGC-3, TTCG-1
EcoRV	GAT†ATC	NIPSR	1AZO	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	U09581	954	-	317	GGCGCCGG-1, GCCGCGG-1, CGGCGC-2, GCCGCG-1, CGCGC-2, GCCGG-1, CGCGG-2, GCGGC-2, GGCCG-2, CGGCC-1, CGGC-3, CCGG-5, GGCG3, CCGC-4, GCCG-3, GGCC-1, CGCG-4
NgoMIV	G†CCGGC	APRG	1FIU	M86915	861	P31032	286	CCGGCG-1, GCCGC-4, GGCG-1, GCCG-3, CGCC-2, CCGC-1, GCGG-2, CGGC-2
PvuII	CAG†CTG	OSALCA	1PVI	AF305615	474	P23657	157	AGCT-2, GCTG-1

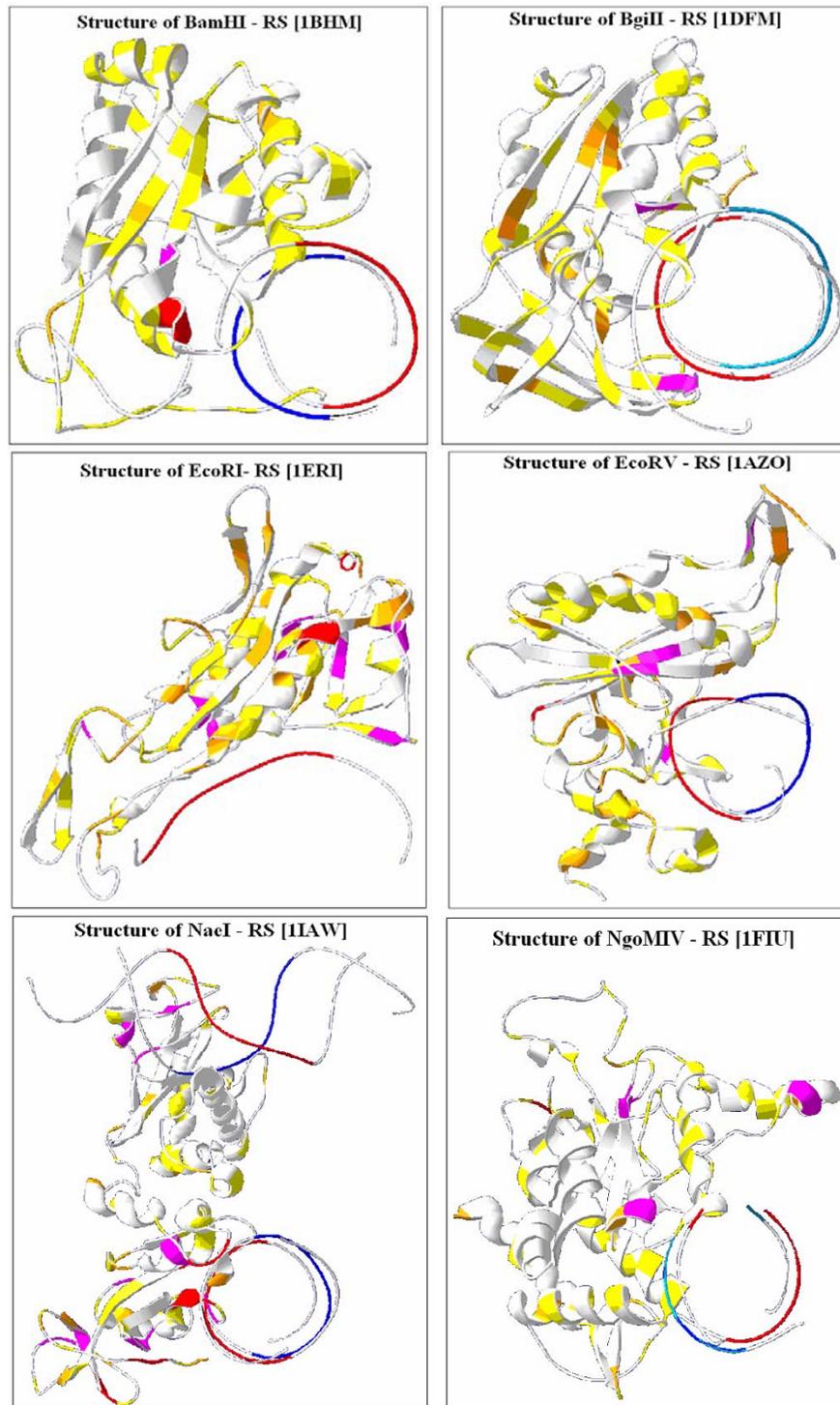AC#: accession number, N.A.: nucleic acid, A.A.: amino acid, †: cut site

Figure 4
The location of RS-like sequences in the 3D structure of REs. The figure shows one subunit of the RE and the dsDNA (the RS). The color code of the ribbon backbone indicates the length of the RS-like strings: yellow = 3, orange = 4, pink = 5, red >= 6. 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).
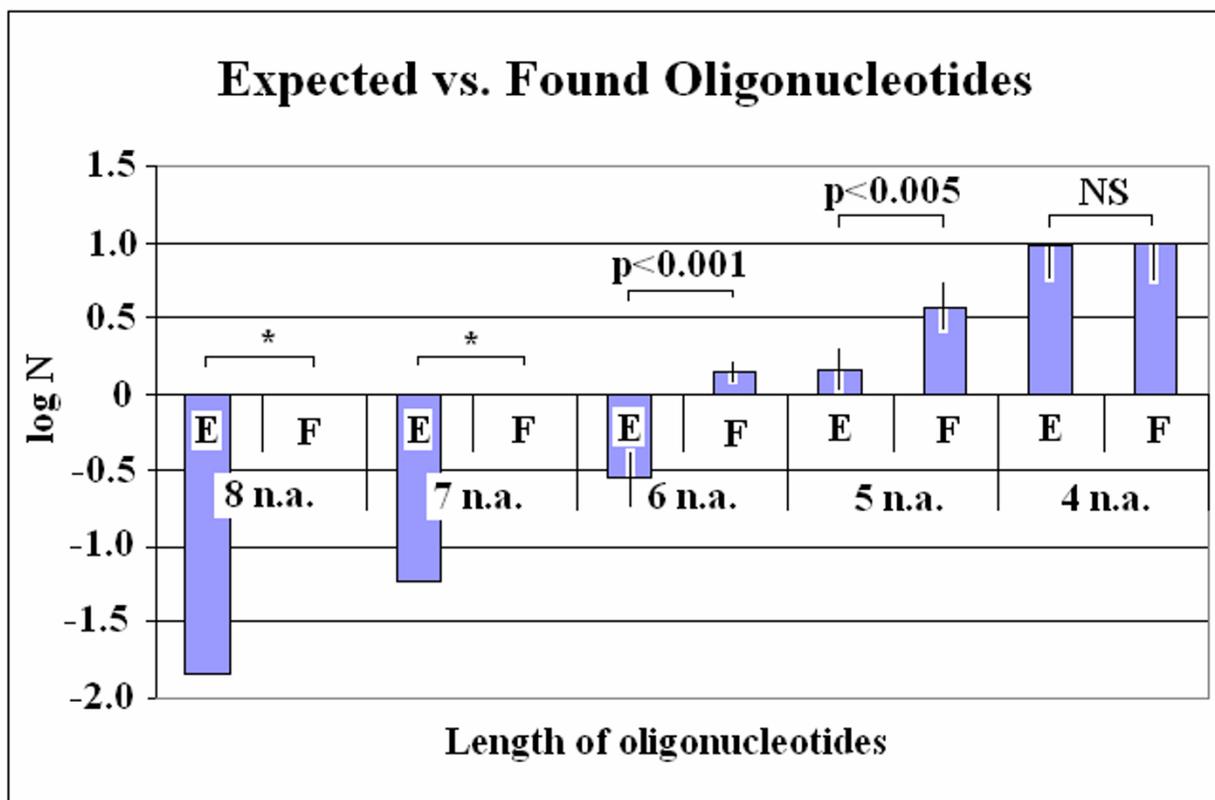
Figure 5
Expected (E) vs. Found (F) RS-like oligonucleotides in the REs. Expected and observed numbers (N) of RS-like nucleotides from 4 to 8 residues long are shown. Statistically significant E – F differences are indicated. NS: not significant, *: single value. For details see the *Results*.

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 phopho-deoxyribosil 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 7). In this example all the four theoretically possible overlappingly translated amino acids of the sequence CGAATT were co-located with the RS (GAATTC).

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 [13]. 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 [14]. 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 [3] propagated the theory of the coevolution of proteins and nucleic acids and argued for a specific stereochemical connection
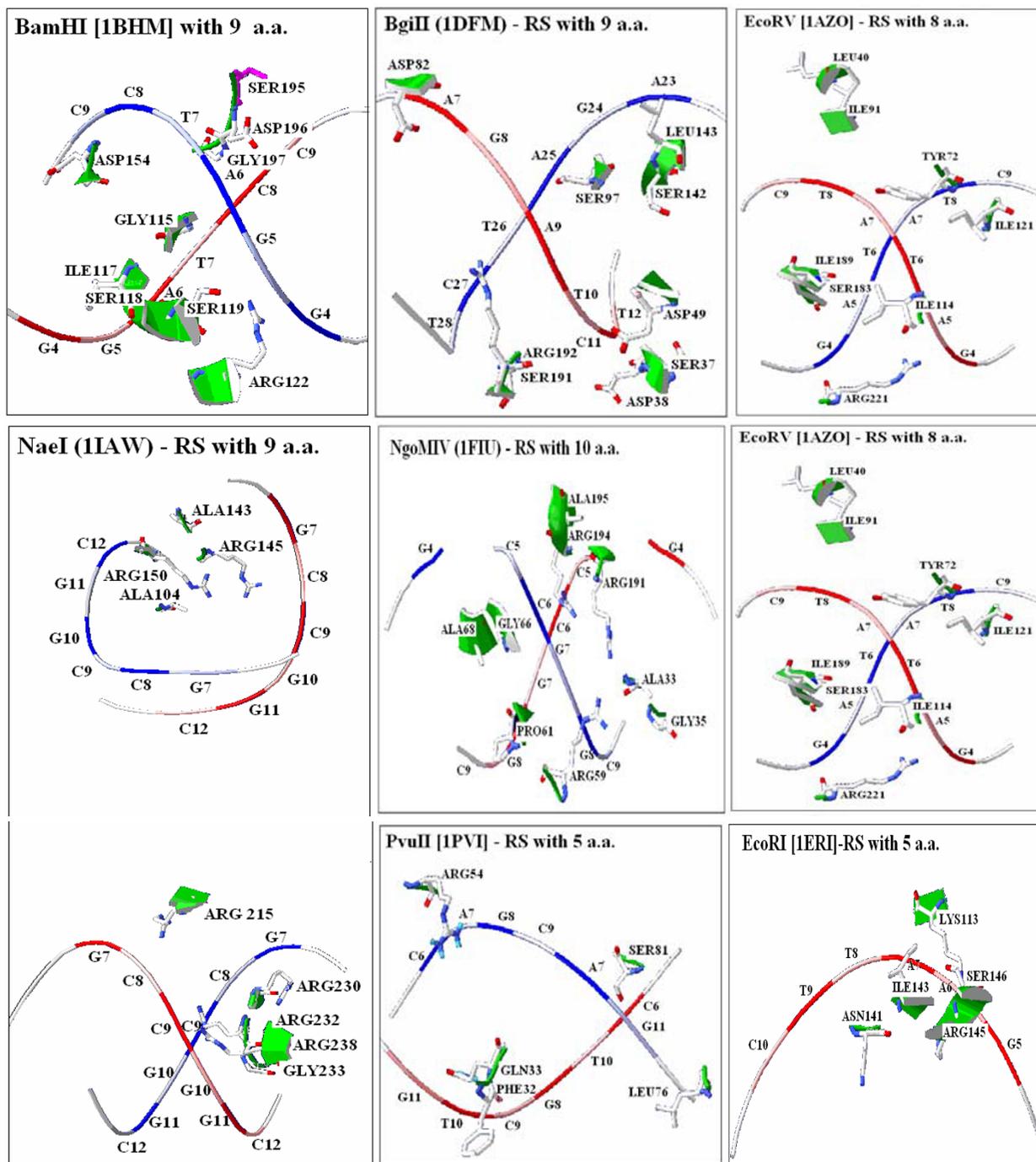
Figure 6
Amino acids in REs coded by RS-like codons and co-located with RSs. The solid spirals indicate the dsDNA; the red/pink and blue/light blue lines are the RSs while the white lines are not RSs. (EcoRI is an exception, there is only a single DNA strand), a.a.: amino acid. The ribbons belonging to the amino acids are green.

Table 2: Codon-amino acid co-location in 7 REs: The proportion of the involved codon residues

ABC%	C-ABC%	AB%	C-AB%	BC%	C-BC%
41.5 +/- 3.6	20.5 +/- 3.1	2.2 +/- 0.4	5 +/- 0.4	18.2 +/- 2.8	18.7 +/- 3.1
	$p < 0.001$		NS		NS

A (1st), B (2nd), C (3rd)-residues in the codon-like sequences of the RSs. C-: nucleic acids in the codon-like sequences of the control (shuffled) RSs. Mean +/- S.E.M., n = 7 (i.e. 7 groups corresponding to the 7 RE). NS: not significant.

Table 3: Codon-amino acid co-location: The shortest atomic distances (Å)

Group	Mean +/- S.E.M. (n)	Significance*
BamHI	6.2_0.9 (9)	NS
BgiII	5.9_0.4 (9)	NS
EcoRI	5.1_0.8 (5)	NS
EcoRV	6.9_0.8 (8)	NS
NaeI	3.9_0.5 (9)	$P < 0.05$
NgoMIV	5.4_0.6 (10)	NS
PvuII	5.0_0.6 (5)	NS
Average	5.5_0.2 (55)	
Hydrophobic	7.3_0.3 (15)	$p < 0.01$
Hydrophilic	5.8_0.3 (19)	NS
Positively charged	3.4_0.2 (15)	$p < 0.001$
Negatively charged	5.3_0.6 (6)	NS

* compared to the average, NS: not significant

between the amino acids and their codons. We succeeded in constructing a "Common Periodic Table of Codons and Amino Acids" [2] 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 was 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 [16]. Codons for alanine, glycine, valine and aspartate have relatively high frequency in the acceptor stems of their respective tRNAs [17]. The tRNAs with complementary anticodons also had some kind of complementarity with their acceptor stems [18]. 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 55 examples where a nucleic acid was co-located with its own codon in such a way that it might indicate a stereospecific interaction.
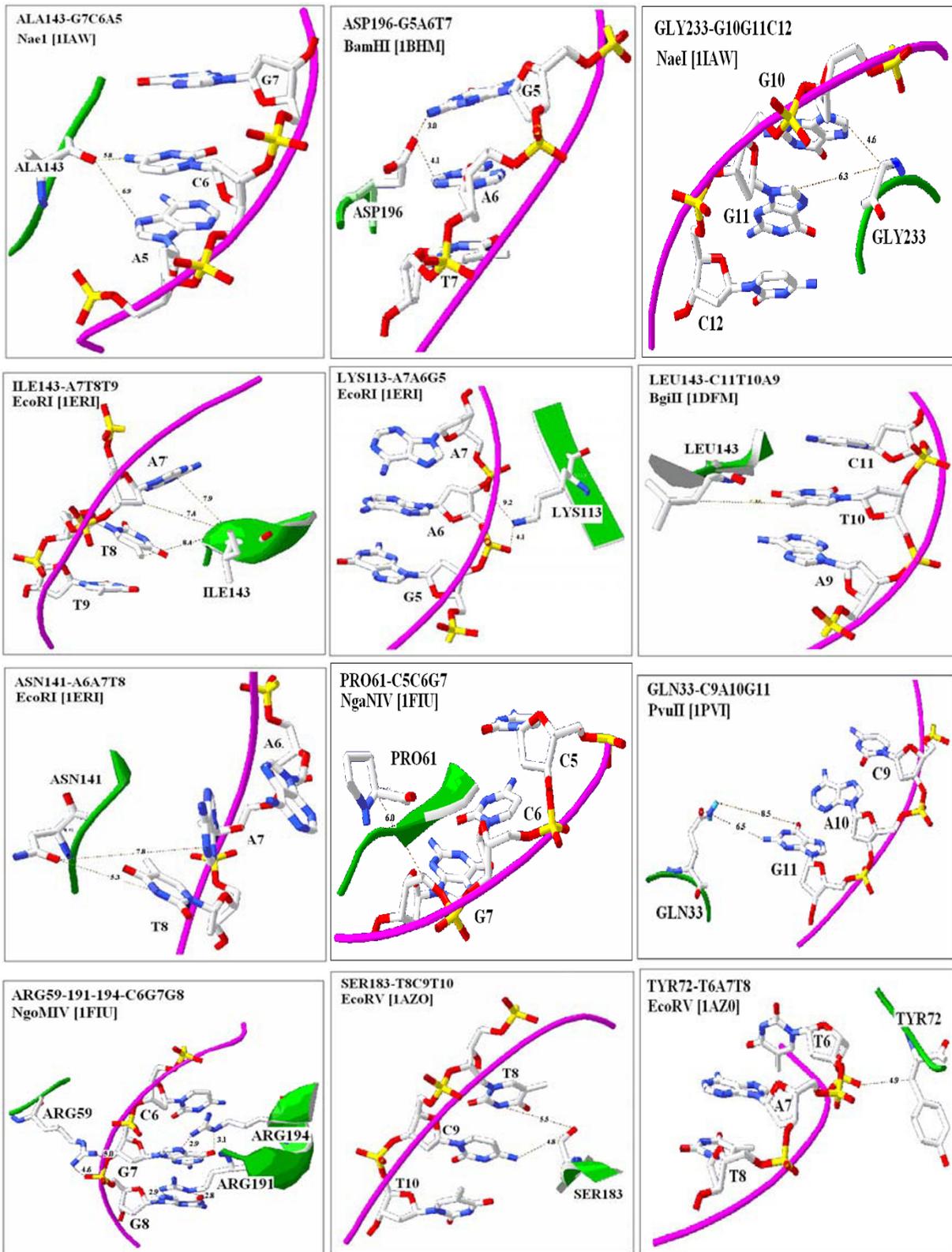
Figure 7
Co-location of codon-like triplets and amino acids in RE-RS complexes. Examples are taken from Figure 6.
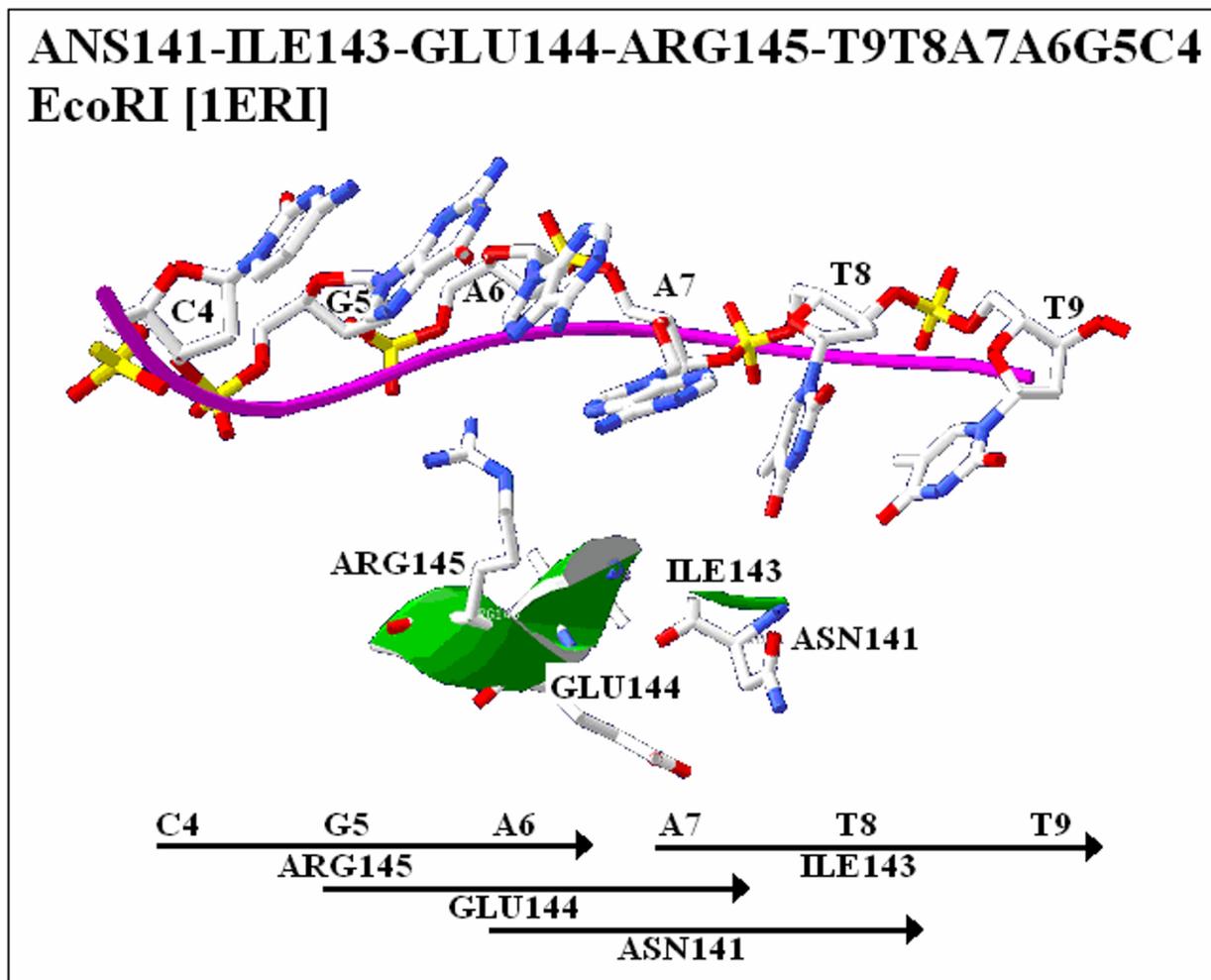
Figure 8
Co-location of codon-like triplets and amino acids in RE-RS complexes. Examples are taken from Figure 6.

In the case of the positively charged amino acids the atoms with opposite partial charge (N, O) were involved and they were close enough to each other to interact. The hydrophobic amino acids were probably too far from their own codon-like triplet in the RS to be in direct interaction with the dsDNA. However the DNA-RE structure is changing during the enzyme reaction and different parts of the protein might be involved during the process.

The number of examples in the presented sample is impressing: 41.5% perfect matches between amino acids and codon-like sequences within 8 Å distance. However 55 examples are still few to categorize the type of interactions and draw a general conclusion.

We are aware of some early model studies [19,20] indicating stereochemical relationship between coding triplets and amino acids, as well as the error in that model building [21]. We don't want to repeat that mistake while searching for a fallen apple close to its tree.

Conclusions

Our previous research, the construction of a Common Periodic Table for Codons and Amino acids [2] already indicated that the amino- and nucleic acids developed in close connection to each other. The present results provide additional evidence to strengthen Woese's hypotheses [3] that there is a stereochemical connection between amino acids and their codons. It was shown that the

recent bioinformatics methods and existing databases provide realistic conditions to study this question.

References

1. Gill G: **Regulation of the initiation of eukaryotic transcription.** *Essays Biochem* 2001, **37:**33-43.
2. Biro JC, Benyo B, Sansom C, Slavecz A, Fordos G, Micsik T, Benyo Z: **A common periodic table of codons and amino acids.** *Biochem Biophys Res Com* 2003, **306:**408-415.
3. Woese CR: **The Molecular Basis for Gene Expression in: The Genetic Code.** *Volume Chapters 6–7. Harper & Row, New York*; 1967:156-160.
4. Roberts RJ, Vincze T, Posfai J, Macelis D: **REBASE – restriction enzymes and methylases.** *Nucleic Acids Research* 2003, **31:**418-420.
5. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-7.
6. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan L, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
7. Berman HM, Westbrook JZ, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28:**235-242.
8. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-80.
9. Clamp M: **Jalview.** 1999 [http://www.ebi.ac.uk/~michele/jalview/].
10. Biro JC: **Overlapping translation of nucleic acids for bioinformatics applications.** *Med Hypotheses* 2003, **60:**654-659.
11. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-Pdb-Viewer: An environment for comparative protein modeling.** *Electrophoresis* 1997, **18:**2714-2723.
12. **Student's t-test** [http://www.physics.csbsju.edu/stats/t-test_NROW_form.html]
13. Biro JC: **Speculation about alternative DNA structures.** *Med Hypotheses* 2003, **61:**86-97.
14. Watson JD, Baker TA, Bell SP, Gann A, Levine M, Losick R: **The structure of DNA and RNA.** in *Molecular biology of the gene Volume chapter 6*. 5th edition. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York*; 2004:1-33. (preprint in 2003)
15. Crick FHC: **The origin of the genetic code.** *J Mol Biol* 1968, **38:**367-379.
16. Schimmel P: **Origin of genetic code: A needle in the haystack of tRNA sequences.** *Proc Natl Acad Sci USA* 1996, **93:**4521-4522.
17. Moller W, Janssen GM: **Statistical evidence for remnants of the primordial code in the acceptor stem of prokaryotic transfer RNA.** *J Mol Evol* 1992, **34:**471-477.
18. Rodin S, Rodin A, Ohno S: **The presence of codon-anticodon pairs in the acceptor stem of tRNAs.** *Proc Natl Acad Sci USA* 1996, **93:**4537-4542.
19. Pelc SR, Welton MGE: **Stereochemical relationship between coding triplets and amino-acids.** *Nature* 1966, **209:**868-870.
20. Welton MGE, Pelc SR: **Specificity of the Stereochemical relationship between ribonucleic acid-triplets and amino-acids.** *Nature* 1966, **209:**870-872.
21. Crick FHC: **An Error in Model Building.** *Nature* 1967, **213:**798.