Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

Biophysical *Journal*
Review

Biophysical Society

Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition

Martin Egli[1,*] and Shuguang Zhang[2]
[1]Department of Biochemistry, Vanderbilt University, School of Medicine, Nashville, Tennessee and [2]Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT Fifty years ago, the first atomic-resolution structure of a nucleic acid double helix, the mini-duplex $(ApU)_2$, revealed details of basepair geometry, stacking, sugar conformation, and backbone torsion angles, thereby superseding earlier models based on x-ray fiber diffraction, including the original DNA double helix proposed by Watson and Crick. Just 3 years later, in 1976, Ned Seeman, John Rosenberg, and Alex Rich leapt from their structures of mini-duplexes and H-bonding motifs between bases in small-molecule structures and transfer RNA to predicting how proteins could sequence specifically recognize double helix nucleic acids. They proposed interactions between amino acid side chains and nucleobases mediated by two hydrogen bonds in the major or minor grooves. One of these, the arginine-guanine pair, emerged as the most favored amino acid-base interaction in experimental structures of protein-nucleic acid complexes determined since 1986. In this brief review we revisit the pioneering work by Seeman et al. and discuss the importance of the arginine-guanine pairing motif.

SIGNIFICANCE Protein-nucleic acid interactions are central to biological information transfer and require sequence-specific recognition. Processing factors, such as nucleases and proteins involved in transcription, rely on precise readout of the information stored in DNA. Packaging and condensation of DNA are less sensitive to base sequence. Remarkably, base-amino acid binding motifs proposed to mediate sequence-specific recognition between DNA and proteins years before the advent of the first experimental protein-DNA complex structure have now been found in thousands of cases. The creative mind envisions concepts with sparse information and correctly predicting DNA readout modes is a case in point. Thinking deeply about DNA early-on led Ned Seeman to propose how proteins read DNA; thinking differently about it led him to create DNA nanotechnology.

INTRODUCTION

Five decades ago, when Apollo 15 astronauts landed on the moon and first rode on its surface in a rover and the Mariner 9 probe was launched toward Mars, the structure of DNA was still defined by a low-resolution model based on fiber diffraction. Single-crystal x-ray crystallography had yielded a host of structures of purines, pyrimidines, and their intermolecular complexes (1), but an atomic-resolution model of the DNA double helix remained elusive. The situation changed between 1971 and 1973, when crystal structures of uridinyl-3′,5′-adenosine phosphate (UpA) in the single-stranded state (2,3) ("DNA UpA movie Aug 1971" https://www.youtube.com/watch?v=PraieqBi048) and then adenosyl-3′,5′-uridine (ApU) paired to itself were determined (4,5).

Ned Seeman, who passed away aged 75 on November 16, 2021, played a key role in the elucidation of the structures of these dimers (Fig. 1) (6). As described by him (7), he arrived in Alex Rich's lab at MIT after obtaining his Ph.D. in small-molecule crystallography at the University of Pittsburgh and a postdoctoral year at Columbia University. At Columbia, he was instrumental in the solution of the first dimer structure, that of UpA. Ned thrived in Alex's lab and proved central to cracking three more dinucleoside structures there, including the one of ApU that revealed Watson-Crick basepairing at high resolution for the first time. In his essay, Ned points out that he had not crystallized any of the dimers, but used his skills in ferreting out structures, and refers to the

Submitted April 13, 2022, and accepted for publication June 10, 2022.
*Correspondence: martin.egli@vanderbilt.edu
Editor: Tamar Schlick.
https://doi.org/10.1016/j.bpj.2022.06.017

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

FIGURE 1 Ned Seeman in his office at New York University in June 2019 (photo credit: Shuguang Zhang). Seeman's office was a treasure trove with DNA origami objects accumulated over the years and many other interesting scientific articles and artifacts. It may appear to be chaotic, but there is order in the chaos. To see this figure in color, go online.

pairing mode seen in the ApU dimer as "a lucky punch." He subsequently tried his hands on an intercalator, 9-aminoacridine, and growing a crystal together with a dinucleoside. However, the structure of the cocrystal failed to show the anticipated intrahelical intercalation (8). In Ned's own words "It left a somewhat bitter taste in my mouth to have so little control over the structure of the crystal, a problem I would ultimately address" (7). Here, we can see the roots of Ned's aspiration to control crystalline structure ("crystal engineering") and use DNA to construct lattices with predictable dimensions and symmetries. His founding of DNA nanotechnology can thus be traced back to the early days of oligonucleotide crystallography, when structures of dimers afforded important insights into the stereochemistry of basepairing and the helical parameters of paired antiparallel strands. Based on these, it was possible to construct models of longer duplexes, as described in a paper by Rosenberg, Seeman, and co-workers (9). Ned Seeman's years at MIT also coincide with another early breakthrough in nucleic acid crystallography, the elucidation of the 3D structure of transfer RNA (tRNA) (10–12). The structure of $tRNA^{Phe}$ at 3 Å yielded insights into many unusual interactions between bases besides the canonical Watson-Crick pairs, including Hoogsteen and wobble pairs as well as diverse base triplets. Thus, by 1975, the geometry of double helix nucleic acids, distinct features of A:U, U:A, G:C, and C:G pairs, and a host of alternative interaction modes between nucleobases were known in quite some detail.

Despite considerable progress in studies directed at the structure of the nucleic acids and the intense interest in determining detailed structures of proteins and enzymes, there was no clear understanding of the interplay between the two key macromolecular species in the mid-1970s. A look at the statistics pages in the Protein Data Bank (http://www.rcsb.org) (13) shows that, by 1976, the coordinates of just 13 protein structures had been deposited. No structure of a nucleic acid-binding protein had been analyzed at that time. In 1977 Alex Rich provided an overview of protein-nucleic interactions with lists of proteins that interact with DNA and proteins that interact with RNA (14). Among the former were proteins involved in DNA replication and repair, packaging, and transcription, as well as nucleases and modifying enzymes. Among the latter were proteins interacting with ribosomal RNA, messenger RNA, and tRNA, as well as polymerizing and repair enzymes and nucleases. Protein-nucleic interactions were anticipated to entail various degrees of specificity; ribosome and nucleosome offer a clear contrast in this respect. Thus, the ribosomal particle involves a very large number of proteins, many of which were expected to be engaged in specific contacts with segments of ribosomal RNA. By comparison, the much smaller nucleosome particle contains just five major proteins and the interactions with double-helical segments of DNA appeared to be largely nonspecific. Conversely, the class of proteins involved in the transcription of DNA, nucleases, and many others rely on sequence-specific recognition of double helical DNA. How such specific contacts could be mediated, the architecture of DNA-binding domains, and whether there was a limited set of motifs used for reading the DNA sequence or a much wider array, remained enigmatic.

Given the lack of information, both with regard to the structure of nucleic acid binding proteins and the nature of the interactions between nucleic acids and proteins, the paper with predictions of "Sequence-specific recognition of double helical nucleic acids by proteins" published by Ned Seeman, John Rosenberg, and Alex Rich in 1976 (15) is nothing short of astonishing (16). In this brief review, we revisit the pioneering work by Seeman et al. and highlight one of the interaction motifs proposed at the time, the major groove contact between arginine and guanine. Since its inception, this motif has been observed in thousands of structures of protein-nucleic acid complexes. The first crystal structure of a protein-DNA complex was reported in 1986 (17). In the complex between *Eco*RI endonuclease and the cognate DNA oligonucleotide d(TCGCGAATTCGCG), arginine 145 is inserted into the major groove of the duplex, but interacts with a phosphate and N7 of adenine (18) (PDB: 1ERI). One of the earliest examples of a guanine-arginine major groove contact in an experimentally determined structure is seen in the complex between the DNA operator and the MAT alpha 2 homeodomain (19) (PDB: 1APL).

BASEPAIR RECOGNITION BY AMINO ACID SIDE CHAINS REQUIRES TWO H-BONDS: THE ORIGINAL MOTIFS

To tackle the problem of sequence-specific recognition of the four basepairs in a double helix, i.e., A:U(T), U(T):A, G:C, and C:G, Seeman et al. first compared different types of basepairs by superimposing them (15). The geometries are those observed in the crystal structures of mini-duplexes

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

and both basepairs were attached to the same ribose moieties in projections along the vertical to the base planes, thereby eliminating the helical twist. This results in four combinations: (1) A:U versus U:A, (2) G:C versus C:G, (3) A:U versus C:G, and (4) A:U versus G:C (Fig. 2). In the illustrations, the wide major groove is at the top and the small minor groove is at the bottom. Potential recognition sites for discriminating between basepairs are marked W in the wide groove and S in the small groove. W1 marks C5 of U (C5m of T) or N7 of a purine. W2 is O4 of U/T or N4 of C, and W3 is N6 of A or O6 of G. W1′, W2′, and W3′ are related to these three by the local dyad that transforms the ribose of the first strand into that of the opposite one. In the small groove, S1 marks either O2 of pyrimidine or N3 of purine, and S2 marks C2 of A and, slightly shifted and almost coinciding with the dyad position, N2 of G (Fig. 2).

Irrespective of the folding motifs proteins might use to interact with nucleic acid duplexes, distinguishing between basepairs involves the grooves as the hydrophobic base stack in the core of the duplex exposes the edges of basepairs at the floor of both grooves. The backbones were thought of as a frame of reference that proteins would use to probe basepairs. To achieve the latter, H-bonding was considered superior to hydrophobic interactions, e.g., stacking, owing to the specificity and directional character of H-bonds. These electrostatic interactions would be mediated by certain protein side chains. To further simplify the problem at hand, only interactions occurring in the plane defined by a single basepair were considered. Thus, a protein side chain contacting both nucleic acid backbone and basepair edge or two adjacent basepairs were ignored to analyze the problem of sequence specificity. This led to the first important question, namely whether it was possible to discriminate between Watson-Crick basepairs depicted in Fig. 2 by single interactions (15)? For example, the W1 and W1′ recognition sites in the outer major groove could provide clear discrimination between alternative pairs shown in Fig. 2 *A–C*. However, a single interaction with W1 or W1′ results in potential ambiguities when the overlay shown in Fig. 2 *D* is considered. Similarly, recognition sites W2/W3 and W2′/W3′ in the central major groove afford discrimination, except in the case of the overlay shown in Fig. 2 *C*. A slight movement of the interacting protein side chain atom vis-à-vis basepair acceptor and donor moieties might result in ambiguities in this case. In the minor groove there are only three recognition sites and a single interaction by a protein side chain with S1 or S1′ in the outer minor groove

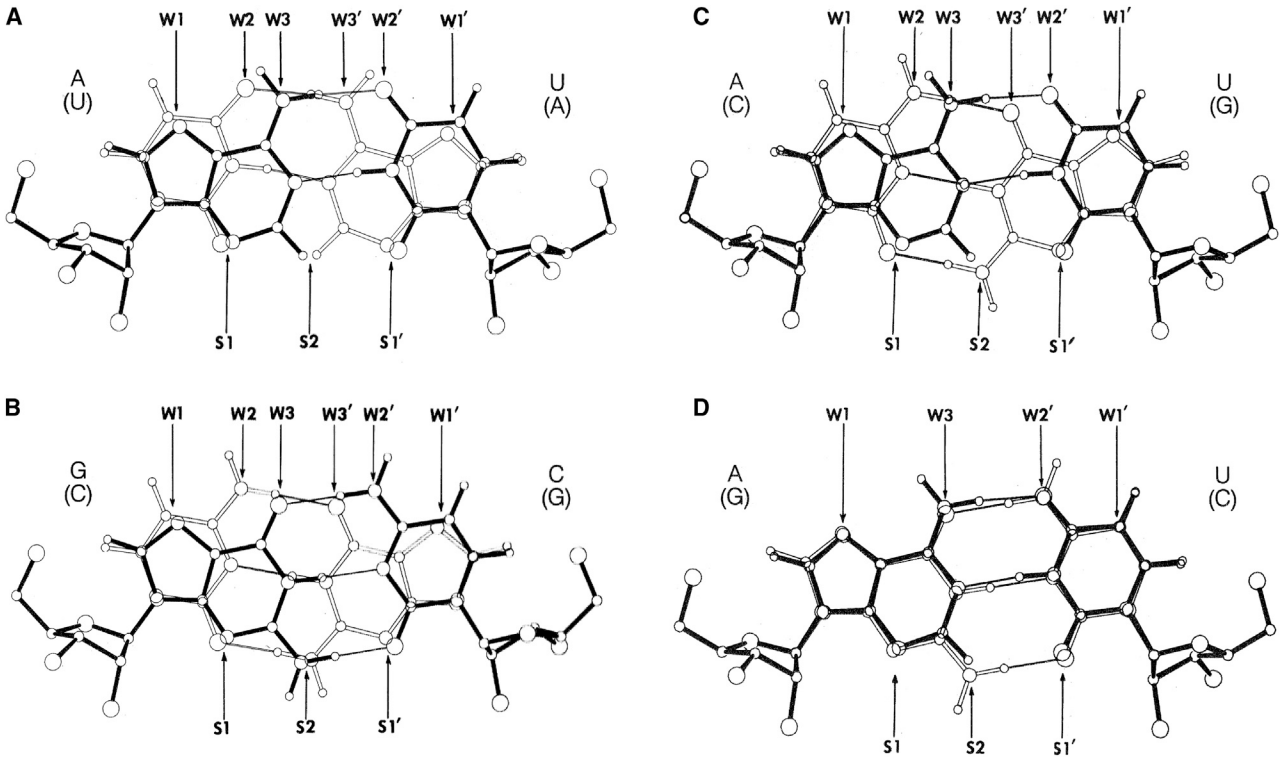

FIGURE 2 Pairwise overlays of A:U and G:C pairs representing four possible comparisons. Top and bottom pairs are shown with solid and open bonds, respectively (*A–D*). Upper letters at the side refer to top bases and lower letters in parentheses refer to bottom bases. Bases are attached to the same ribose and the twofold rotation axis relating sugar moieties runs along the vertical in the plane of the paper, roughly along the arrow marked S2 (*A–C*). W1-3 and W1′-3′ refer to recognition sites in the major or wide groove of the duplex, and S1, S1′, and S2 refer to recognition sites in the minor or small groove. Reproduced with permission from: Seeman, N.C., Rosenberg, J.M., Rich, A. 1976. Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. USA, 73:804–808.

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

would be unable to discriminate between overlaid pairs shown in Fig. 2 *D*. Using S2, discriminating between A:U/U:A (Fig. 2 *A*) or G:C/C:G (Fig. 2 *B*; considering the possibility of a slight movement of the interacting protein side chain in the minor groove) would not be possible. Therefore, this simple analysis that considers a couple of recognition sites for four combinations of overlaid basepairs demonstrated that a single interaction is not suitable for sequence-specific discrimination between all of them with sufficient precision.

Unlike a single H-bond formed between a protein side chain and the edge of a basepair in either groove, pairs of H-bonds can establish unique interactions that allow sequence-specific recognition of double helix nucleic acids by proteins. The motifs proposed by Seeman et al. were inspired by different types of base triplets, e.g., U:A·U and C:G·G, seen in the structures of polynucleotides, complexes between nucleobases, or $tRNA^{Phe}$ (15). Thus, guanine in the C:G pair could be recognized by arginine using its guanidino moiety to form H-bonds to O6 and N7 from G in the major groove (Fig. 3 *A*). Adenine in the U:A pair could be recognized by asparagine or glutamine using their amide moiety to form H-bonds to N6 and N7 from A in the major groove (Fig. 3 *B*). Similarly, guanine could be recognized by either asparagine or glutamine forming H-bonds to N2 and N3 from G in the minor groove (Fig. 3 *C*).

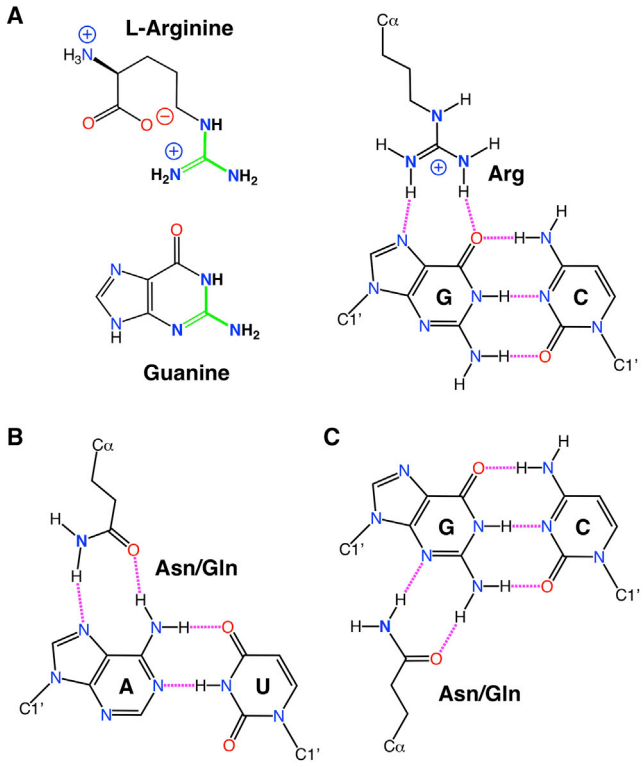

FIGURE 3 Pairwise H-bonding motifs between amino acids and nucleobases proposed by Seeman et al. in 1976 for mediating sequence-specific recognition of basepairs by proteins (15). (*A*) Interaction between arginine and G:C in the major groove (right) and comparison between guanine and arginine with guanidino moieties highlighted with green bonds (left). (*B*) Interaction between asparagine or glutamine and A:U in the major groove. (*C*) Interaction between asparagine or glutamine and G:C in the minor groove. One can envision that an additional carboxyl moiety (Asp/Glu) could bridge N4 of C and the arginine amino group in the Arg ··· G:C motif depicted in (*A*). Similarly, an additional amino moiety (Lys) could bridge O4 of U and the asparagine (glutamine) amide oxygen atom in the Asn (Gln) ··· A:U motif depicted in (*B*). To see this figure in color, go online.

ARGININE-G PAIRS IN PROTEIN-DNA CRYSTAL STRUCTURES

Some 15 years after the publication of the sequence-specific recognition of nucleic acid double helices by proteins in the *Proceedings* (15), a large enough number of protein-DNA complex crystal structures had been determined to take stock and examine the experimental evidence in regard to predicted motifs (20). Indeed, these structures confirmed direct interactions between amino acids and nucleobases of the types suggested years earlier. Thus, the DNA complexes of *Eco*RI, Trp repressor, λ cro repressor, mouse zinc finger protein, and glucocorticoid receptor all revealed arginine to guanine H-bonding. The DNA complexes of λ repressor, 434 repressor, 434 cro repressor, homeodomain, and λ cro repressor showed either glutamine or asparagine side chains forming H-bond pairs with adenine. The experimental structures at long last shed light on the diverse scaffolds that proteins erect to interact with DNA, including the helix-loop-helix binding motif in the major groove and repeat recognition modules, such as zinc fingers, along with the role of water in mediating sequence-specific contacts, e.g., in the Trp-repressor-DNA complex (21). The structural diversity of DNA binding proteins has been described in detail (22–24). Here, we focus on the ubiquitous occurrence of interactions between arginine and the Hoogsteen edge of guanine.

The recognition motif in the major groove predicted by Seeman et al., i.e., a kind of "head-on" approach of arginine to use the amines of its guanidino moiety for H-bonds with the guanine O6 and N7 acceptors (Fig. 3 *A*), is found in the majority of all protein-DNA complexes studied to date. We can refer to this mode of interaction as canonical (recognition motif) and examples are depicted in Fig. 4. Importantly, this interaction is seen for proteins and enzymes that use very different means to bind to and interact with DNA and exert a diverse set of functions (22–24). With regard to the DNA binding motif, the examples selected for Fig. 4 include α-helices (helix-turn-helix), antiparallel β-strands, zinc finger modules, extended coils, and loop regions. In every case, the arginine guanidino moiety approaches G more or less within the plane of the guanine base and establishes two H-bonds to the edge available in the major groove. In their paper, Seeman et al. state that they arbitrarily used the two guanidino amino groups of arginine to contact guanine, although it was clear that one amino and one imino

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

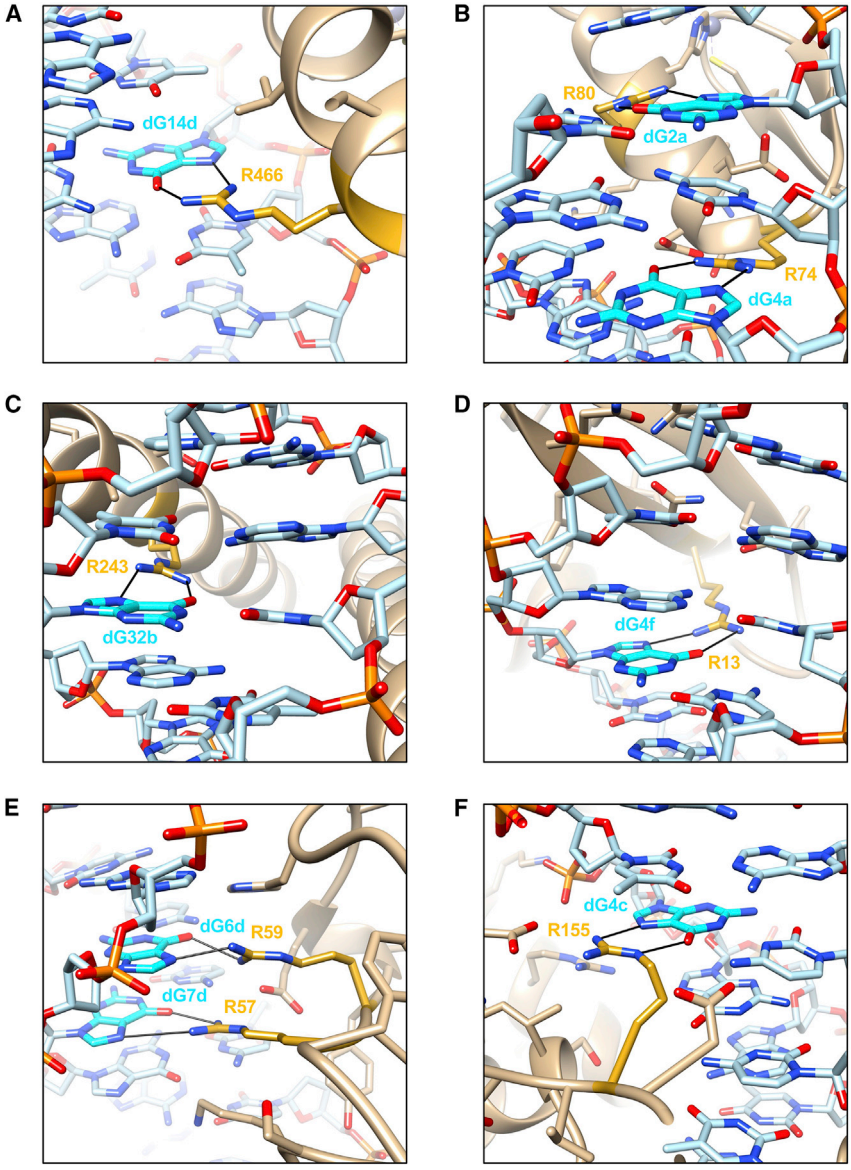

FIGURE 4 Examples of arginine-guanine pairs in crystal structures of protein-DNA complexes. (*A*) Glucocorticoid receptor, PDB: 1R4O; α-helix inserted into the major groove (25). (*B*) Zinc finger Zif268, PDB: 1ZAA; zinc finger module visible in the background with Zn^{2+} shown as a gray sphere (26). (*C*) GCN4 basic region leucine zipper, PDB: 1YSA; extended α-helices in the major groove (27). (*D*) Arc repressor, PDB: 1PAR; recognition of DNA by β-sheets (28). (*E*) NF-κB p50 Rel-homology region, PDB: 1SVC; β-barrel domains that grip DNA in the major groove (29). (*F*) *Bam*HI endonuclease, PDB: 1BHM; DNA binding in a cleft with the enzyme contacting several basepairs mainly in the major groove (30). To see this figure in color, go online.

group could serve as H-bond donors to interact with G (15). An example of the latter type is depicted in Fig. 4 *F*. Arginine approaches the base edge from one side of the major groove to form a pair of H-bonds and also place its positive charge in close vicinity of the phosphate backbone.

An analysis of amino acid-nucleobase interactions in 129 protein-DNA complexes demonstrated that the arginine-guanine pair is the most common motif—98 observations (31). By comparison, there were 218 arginine-phosphate salt bridges, and lysine was interacting with guanine and phosphate in 30 and 109 cases, respectively. Remarkably, the runner-up in terms of amino acid-base interactions was adenine forming H-bonds with asparagine or glutamine (34 observations), a recognition motif also predicted by Seeman et al. (Fig. 3, *B* and *C*). A more recent investigation of sequence-specific recognition of DNA by proteins coupled energetic favorability and geometric preferences and included a much larger set of experimental structures (32). A total of 1584 crystal structures of protein-DNA complexes with a resolution better than 2.5 Å and an R-factor no higher than 25% were analyzed in regard to amino acid-basepairing motifs of optimal stability and geometry. For a single guanine, arginine interacting with the Hoogsteen edge (O6/N7) and aspartate interacting with the Watson-Crick edge (N1/N2) via two H-bonds were the winners. Once again, for guanine engaged in a basepair, the analysis confirms that the arginine-guanine interaction as proposed by Seeman et al. is the favored recognition motif. In 1976, the issue of protein-DNA recognition necessarily remained limited to specific amino acid-basepairing motifs (direct readout or sequence recognition). A better understanding of other contributions to recognition, i.e., protein-induced

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

changes in the local conformation of the DNA double helix (indirect readout) had to await the experimental determination of structures of protein-DNA complexes. Thus, despite the ubiquitous nature of the arginine-guanine pair, there are certainly structures of protein-DNA complexes that do not feature arginine H-bonding to the major groove edge of G. One prominent example is the TATA-box binding protein that straddles the minor groove and inserts phenylalanines between basepairs (33,34), thereby inducing a DNA conformation that resembles the A-form (TA-DNA) (35).

Arginine cannot just establish multiple H-bonds and salt bridges, but its guanidino moiety can engage in stacking interactions and the methylene moieties of the long side chain are used for hydrophobic contacts. This chemical versatility—electrostatics boosted by a positive charge, dispersive attractive interactions and hydrophobics—explains why arginine interactions with DNA are necessarily more complex than the pair of H-bonds to the Hoogsteen edge of G that serves sequence-specific recognition (Fig. 4). Owing to the positive charge, nonspecific arginine interactions with the phosphate backbone are more common than interactions with bases that serve recognition in the major groove (31). The crystal structure of the nucleosome core particle shows multiple arginines that are inserted into the narrow DNA minor groove (36), consistent with the relatively strong negative electrostatic surface potential at such sites. Some examples of noncanonical arginine or guanidino moiety interactions with nucleobases in the major groove or at DNA polymerase active sites are depicted in Fig. 5. At the active site of the error-bypass DNA polymerase η, arginine 61 engages in the more or less canonical dual H-bond interaction with the G of the incoming nucleotide, but also bridges the O6 keto oxygens of incoming dGTP and template 8-oxo-dG, and forms a cation $\cdots\pi$ interaction with the 3′-terminal dT of the DNA primer strand (Fig. 5 *A*). The combination of H-bonding and stacking by arginine involving a basepair step, here dT and dG(TP) before phosphodiester bond formation, is actually quite common. A similar thymine-arginine-guanine triplet is observed in the crystal structure of the sporulation regulator Ndt80 in complex with DNA, where tandem arginines pull out thymines from underneath the 3′-adjacent guanines at respective d(TpG) steps to facilitate stacking interactions with the former, thereby exploiting sequence-dependent DNA structural malleability (37). Other cases of coupled H-bonding and stacking interactions by arginine are seen in the structure of the catabolite gene activator protein-DNA complex, where arginines from two α-helices are inserted into adjacent major grooves and engage in stacking and H-bonding with dT and 3′-neighboring dG, respectively, thus contributing to the induction of an overall 90° bend into the DNA duplex (38). Arginine often teams up with other amino acids to interact with base edges in the major groove of DNA. One example is shown in Fig. 5 *B*, with arginine forming just a single H-bond to O6 of guanine, but then establishing three additional H-bonds to a neighboring asparagine and cytosine from the adjacent basepair and opposite strand. The classic tandem H-bonding motif between arginine and guanine also occurs in the major groove of RNA. However, the major groove in the canonical A-form duplex is too narrow and such interactions require the groove to be pried open by bulges or other secondary/tertiary structural motifs to provide access to the Hoogsteen edge of G. In the case of the peptide-RNA complex depicted in Fig. 5 *C*, the arginine-guanine interaction occurs at the end of the major groove that is more open, thanks to the adjacent G-quadruplex region. Finally, in the possibly most radical use of the guanidino moiety of arginine, the side chain displaces a lesioned guanine at the active site of Rev1 error-bypass DNA polymerase. There, it mimics undamaged template G (with which it shares the guanidino moiety, Fig. 3 *A*) to correctly code for incoming dCTP (H-bonds to acceptor atoms N3 and O4, Fig. 5 *D*).

ARGININE FORKS ARE A PROMINENT MOTIF IN PROTEIN-RNA RECOGNITION

The prevalence of the arginine-guanine pair is not limited to DNA-protein interactions; the pair is also an important motif in RNA binding and recognition by proteins. A database that categorized all amino acid-nucleotide interactions in nucleic acid-protein structures deposited in the Protein Data Bank found that such contacts involving arginine were the most common in both DNA and RNA complexes (42). Moreover, guanine was the favored partner of arginine in DNA-protein complexes. In RNA-protein complexes, guanine and cytosine were preferred by arginine relative to adenine and uracil. A subsequent analysis of binding pairs in protein-nucleic acid interactions (database of binding pairs) confirmed the special status of arginine in establishing contacts to DNA and RNA residues (43). However, neither on-line database appears to be accessible anymore as of 2022. A previous study of the frequency of RNA base-amino acid interactions in structures of complexes found that the arginine-guanine pair was the dominant major groove motif and that this preference was correlated with favorable energetics (charge, H-bonding) of the interaction rather than RNA structure (44). An example of an interaction between arginine and the Hoogsteen edge of G that couples binding and catalysis is constituted by arginine 144 and G72 in the complex between *E. coli* prolyl-tRNA synthetase and $tRNA^{Pro}$ (45).

A widespread interaction motif in RNA-protein complexes is the so-called arginine fork (46). In the classical model, the guanidino moiety of the side chain interacts with four nonbridging phosphate oxygens from adjacent phosphates, whereby the latter do not have to be intrastrand or intrahelical. The fork comes in many flavors in the RNA major groove, e.g., arginine bridging 1) two phosphates, 2) a single phosphate and guanine, 3) two phosphates and

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

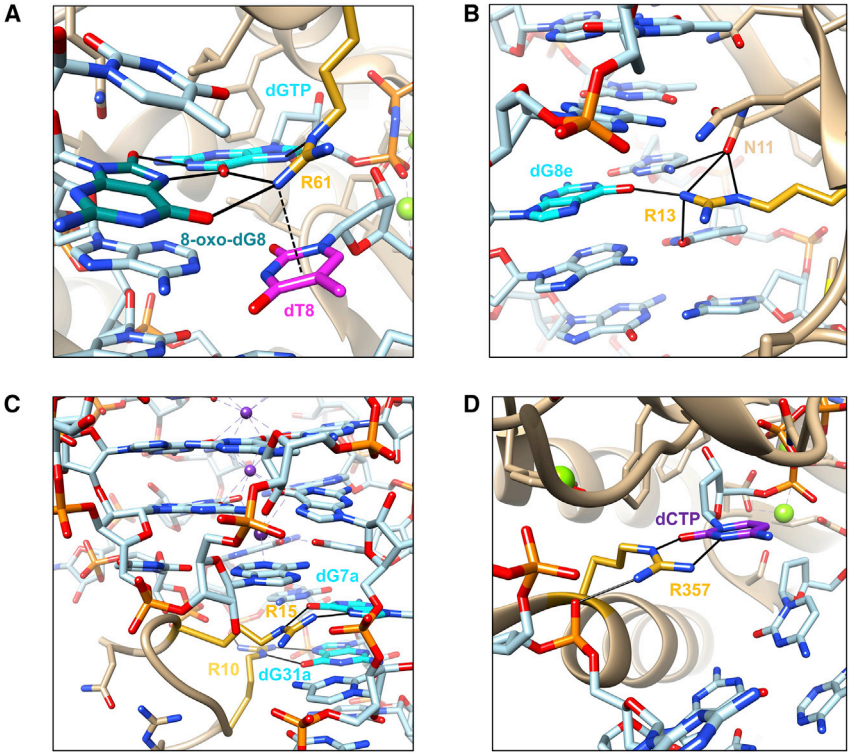

FIGURE 5 Examples of noncanonical arginine-/guanidino moiety-nucleobase interactions in crystal structures of protein-DNA and -RNA complexes. (*A*) Trans-lesion-synthesis Y-family DNA polymerase η, PDB: 4O3Q; view of the polymerase active site, with the enzyme inserting dGTP opposite 8-oxo-dG (39). (*B*) Arc repressor, PDB: 1PAR; arginine teaming up with asparagine and establishing H-bonds to G and T on the opposite strand (28). (*C*) Fragile X mental retardation protein (FMRP), a regulatory RNA binding protein, PDB: 5DEA; the protein uses a β-turn in an arginine-glycine-rich motif to interact with Gs in the major groove adjacent to the G-quadruplex (40). (*D*) Y-Family Rev1 DNA polymerase, PDB: 3GQC; instead of an adducted G from the template (evicted from the active site), an arginine is inserted and uses its guanidino moiety to code for incoming dCTP (41). To see this figure in color, go online.

guanine, and 4) a single phosphate and guanine, coupled with stacking on a second base, and so forth. An example of an arginine fork in the structure of the large subunit of the ribosome is depicted in Fig. 6.

Complexes of amino acid-binding RNA aptamers with their targets offer another opportunity to statistically analyze interactions between side chains and nucleobases. In the case of arginine RNA aptamers, such studies provided support for an intrinsic affinity between the amino acid and its codons (48), perhaps indicating that the genetic code may have a chemical rather than an adaptive basis. Thus, binding sites of aptamers display a strong purine bias (78% of bases at the binding site in the case of arginine aptamers). Furthermore, only arginine aptamers show an overrepresentation of the arginine set of codons (CGN and AGR, where N is any nucleotide and R represents A or G), and only arginine codons are overrepresented in such aptamers. Moreover, arginine codons bind the amino acid better than any other set of codons. Hence, there is a strong association between the arginine codon classes and regions of RNA molecules that interact with the amino acid. As expected, guanine is involved in binding with its major groove edge, among other contributions, and it appears that arginine may have captured the codons for which is possesses the highest affinity. Crystal structures of guanidinium riboswitches offer an interesting perspective in this regard. In the type I riboswitch aptamer structure, recognition of the guanidinium cation involves no fewer than five Gs that bind the ligand via H-bonding, and ionic and cation-π interactions (49). In the type II riboswitch aptamer structure, the guanidinium cation binds within the conserved ACGR tetraloop by H-bonding to the Hoogsteen edge of guanine and forming ionic interactions to three phosphate groups (reminiscent of an arginine fork), as well as engaging in a cation-π interaction with a second G (50). In summary, arginine provides both superior binding strength, which is dominated by the electrostatic contribution in its association with guanine, and versatile recognition strategies that involve H-bonding, ionic interactions, and/or cation-π stacking. Arginine's

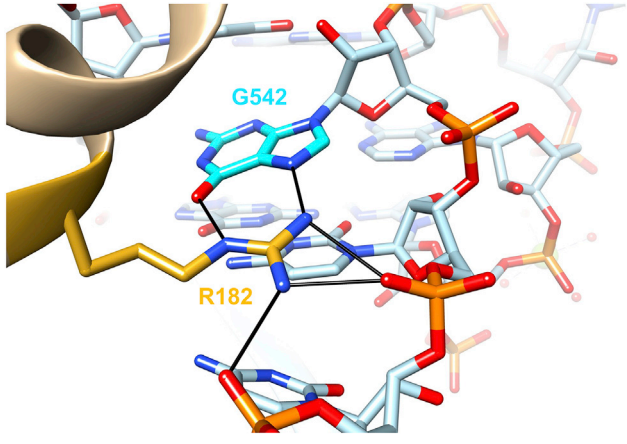

FIGURE 6 Arginine fork in the crystal structure of the large ribosomal subunit: L4 protein bound to 23S rRNA, PDB: 1JJ2 (47). To see this figure in color, go online.

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

advantage over other amino acids—more in the realm of binding strength than recognition in this case—is also evident from the superior ability of arginine peptides to condense DNA (compared with lysine peptides) (51).

GUANIDINO G-CLAMP: A SYNTHETIC CYTOSINE ANALOG THAT FORMS FIVE H-BONDS TO G

The arginine-guanine recognition pair in the major groove first proposed by Seeman et al., and since found in the vast majority of structures of nucleic acid-protein complexes, also inspired the synthesis of a cytosine analog, the guanidino G-clamp, which can potentially form five H-bonds with guanosine (52). The tricyclic 9-(2-guanidino-ethoxy)-phenoxazine moiety features a cytosine core that establishes standard Watson-Crick H-bonds to G and a tethered guanidino group that was intended to form two more H-bonds with the Hoogsteen edge of G. Indeed, an atomic-resolution crystal structure of a modified DNA oligonucleotide confirmed the formation of five H-bonds between the guanidino G-clamp and G (Fig. 7). UV melting assays with DNAs containing a single G-clamp analog paired opposite RNA target strands showed an increase in T_m of up to 16°C compared with the native counterparts. The dramatic gain in stability is the result of the additional H-bonds in the G-clamp:G pair compared with a standard G:C pair as well as improved stacking by the phenoxazine moiety, the presence of a positive charge in the center of the negatively polarized major groove, and an extensive water network that links the guanidino moiety to phosphates from the opposite strand (53).

IN MEMORIAM NED SEEMAN

In this review we revisit early feats in Ned Seeman's scientific career, the determination of several nucleic acid mini-duplex crystal structures and the subsequent prediction of amino acid-base interactions that mediate sequence-specific recognition of double helix nucleic acids by proteins. In the context of the proposed interaction motifs that foreshadowed patterns of direct DNA readout, a quote attributed to Albert Einstein comes to mind: "Imagination is more important than knowledge" (16). Over the last 40 years, Ned's predicted amino acid-nucleobase pairings in the major and minor groove, including the arginine-guanine pair, have been observed in experimentally determined structures of protein-nucleic acid complexes in thousands of cases. Remarkably, the very first structures of protein-nucleic acid complexes that were deposited in the Protein Data Bank (https://www.rcsb.org/stats/growth/growth-protein-na-complex) exhibited amino acid-base interactions proposed by Ned. In the crystal structure of the 434 repressor-DNA complex (PDB: 2OR1, release date September 5, 1989), glutamine forms two H-bonds with the N6/N7 edge of adenine (54). In the x-ray fiber structure of tobacco mosaic virus (PDB: 2TMV, release date January 9, 1989), the coat protein uses arginine to interact with O6 of guanine from the (single-stranded) trinucleotide GAA (55). Ned's take of protein-DNA recognition was both pioneering and incredibly imaginative. It reminds one of a statement in a recent issue of *The Atlantic* magazine on artificial intelligence: "While the way to wisdom leads through knowledge, there is no path to wisdom from information" (56). Well, not true in Ned's case—he jumped directly from very little information to wisdom!

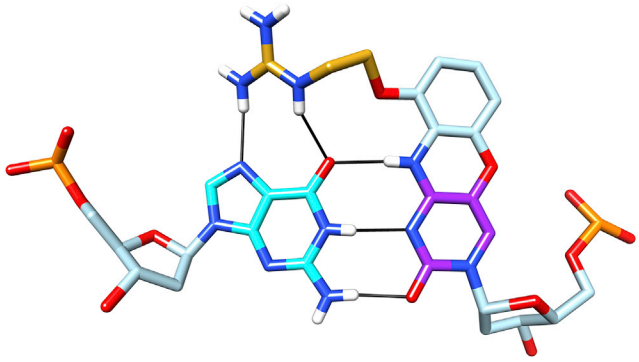

FIGURE 7 The cytidine analog guanidino G-clamp paired to G in the crystal structure of a modified DNA decamer visualized at 1 Å resolution, PDB: 1KGK (52). The color code is the same as that used in Figs. 4 and 5, i.e., guanine, arginine, and cytosine carbon atoms are highlighted in cyan, goldenrod, and purple, respectively. Hydrogen atoms are white and H-bonds are thin solid lines. To see this figure in color, go online.

It is not surprising that Ned saw a different application of DNA when he founded the field of DNA nanotechnology, at a time when most everybody talked about sequencing, genomics, and RNA structure and function. This is yet another example of Ned's boundless imagination, creativity, and curiosity. It was Francis Crick who remarked: "One could not be dedicated to anything unless one believed in it passionately." Ned's passion of using DNA as a construction material paved the road to DNA origami, which had a transformative impact on science and molecular art.

One of the authors, Shuguang Zhang, first met Ned at the "Conversation in Biomolecular Stereodynamics III" meeting in June 1983 at the State University of New York at Albany where he was a professor. In those days, the Who's Who in structural biology and biophysics attended the "Conversations," including Nobel laureate Dorothy Hodgkin and future Nobel laureates Tom Steitz and Martin Karplus. Alex Rich who was Ned's and our postdoctoral mentor regularly gave the opening lectures. Ned loved to have parties and his insatiable appetite to converse with colleagues and friends was legendary. At one of these, most speakers at the conference came to Ned's small apartment, countless bottles of beer and wine were opened, and so many attended that we literally had to stand face to face with not an inch of space to move. Ned's parties were something everyone looked forward to.

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

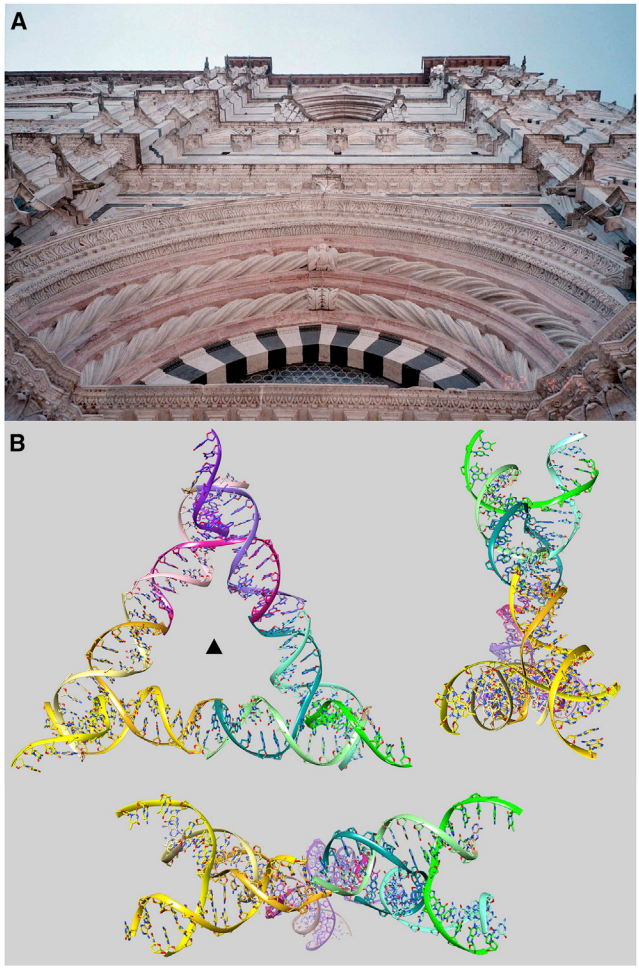

FIGURE 8 Science in art and art in science.
(*A*) Left- and right-handed double helices with major and minor grooves above the portal of the Siena Cathedral (Duomo di Siena) located in the center of the homonymous square in Siena, Italy (photo credit: Shuguang Zhang). Groundbreaking took place in 1196 and the Duomo was completed in 1338, although designs were added over six centuries. The curved double helix at the top is left-handed and the one at the bottom starts out left-handed (from the left) and flips to right-handed behind leaves in the center ("B-Z junction"). (*B*) Crystal structure of a self-assembled tensegrity triangle in space group *R*3, and consisting of the four DNA oligonucleotides d(GAGC AGCCCGTACTCG), d(pCCGAGTACGACGACAAG), d(TCTGATGAGG CTGC), and d(pGCTTGTCGTTCATCA); PDB: 3UBI (58,59). DNA oligos in the three adjacent asymmetric units are colored in shades of yellow, green, and purple. The DNA triangle viewed (counterclockwise from top left): along the threefold rotation axis (*solid triangle*), normal to the yellow/green side of the triangle after a 90° rotation around the horizontal, and along the yellow/purple side of the triangle. To see this figure in color, go online.

At the time of that first meeting in Albany, Ned had already published his seminal theoretical paper on "Nucleic acid junctions and lattices" (57). But Ned was unhappy in Albany where he felt he could not bounce ideas off colleagues. This was in sharp contrast to his productive time at MIT where his original ideas had made an early impact. However, Ned flourished soon after the move to NYU; it was as if an intellectual repressor had been lifted. Ned's detailed knowledge of DNA structure and pairing combined with imagination and interests beyond conventional science as well as an infectious enthusiasm had opened the door to a new field. Soon, many around the world jumped into the DNA origami field and created self-assembled DNA nanoart, including the letters of the alphabet, numbers and diverse symbols, smiley face and portraits of Lincoln and Mona Lisa, thereby following Leonardo Da Vinci's wisdom: "Study the science of art and the art of science" (Fig. 8). One of us (S.Z.) has been teaching Ned's scientific nanoscale art in his Molecular Architecture and Design course at MIT since 2000. Even with Ned now gone, we will thus meet him and his science annually.

Ned was not particularly religious and perhaps he did not really care about the ancient Greek notion of the soul. However, if there were a soul, it's believed to be a person's unique identity that is being continuously remembered and celebrated for years to come, as are Archimedes, Socrates, Pythagoras, Leonardo Da Vinci, Galileo Galilei, Isaac Newton, Wolfgang Amadeus Mozart, Ludwig van Beethoven, Albert Einstein, Francis Crick, Alex Rich, and Ned Seeman.

AUTHOR CONTRIBUTIONS

Both authors wrote and illustrated this review.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Voet, D., and A. Rich. 1970. The crystal structures of purines, pyrimidines and their intermolecular complexes. *Prog. Nucleic Acid Res. Mol. Biol.* 10:183–265.
2. Seeman, N. C., J. L. Sussman, …, S.-H. Kim. 1971. Nucleic acid conformation: crystal structure of a naturally occurring dinucleoside phosphate (UpA). *Nat. New. Biol.* 233:90–92.
3. Sussman, J. L., N. C. Seeman, …, H. M. Berman. 1972. The crystal structure of a naturally occurring dinucleotide phosphate uridylyl-3′,5′-adenosine phosphate. Models for RNA chain folding. *J. Mol. Biol.* 66:403–421.
4. Rosenberg, J. M., N. C. Seeman, …, A. Rich. 1973. Double helix at atomic resolution. *Nature.* 243:150–154.
5. Seeman, N. C., J. M. Rosenberg, …, A. Rich. 1976. RNA double helical fragments at atomic resolution: the crystal and molecular structure of sodium adenylyl-3′,5′-uridine hexahydrate. *J. Mol. Biol.* 104:109–144.
6. Ball, P. 2021. Obituary Ned Seeman (1945–2021). *Nature.* 600:605.
7. Seeman, N. C. 2018. Five years with Alex Rich (1972–1977). *In* The Excitement of Discovery: Selected Papers of Alexander Rich. A Tribute to Alexander Rich. S. Zhang, ed. World Scientific, p. 538. Series in Structural Biology.
8. Seeman, N. C., R. O. Day, and A. Rich. 1975. Nucleic acid-mutagen interactions: crystal structure of adenylyl-3′,5′-uridine plus 9-aminoacridine. *Nature.* 253:324–326.
9. Rosenberg, J. M., N. C. Seeman, …, A. Rich. 1976. RNA double helices generated from crystal structures of double helical dinucleoside phosphates. *Biochem. Biophys. Res. Commun.* 69:979–987.

Please cite this article in press as: Egli and Zhang, Ned Seeman and the prediction of amino acid-basepair motifs mediating protein-nucleic acid recognition, Biophysical Journal (2022), https://doi.org/10.1016/j.bpj.2022.06.017

10. Kim, S. H., F. L. Suddath, …, A. Rich. 1974. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science.* 185:435–440.
11. Suddath, F. L., G. J. Quigley, …, A. Rich. 1974. Three-dimensional structure of yeast phenylalanine transfer RNA at 3 Å resolution. *Nature.* 248:20–24.
12. Rich, A., and S. H. Kim. 1978. The three-dimensional structure of transfer RNA. *Sci. Am.* 238:52–62.
13. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
14. Rich, A. 1977. An overview of protein-nucleic acid interactions. *In* Nucleic Acid-Protein Recognition. Academic Press Inc., pp. 3–11.
15. Seeman, N. C., J. M. Rosenberg, and A. Rich. 1976. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA.* 73:804–808.
16. Egli, M., and S. Zhang. 2022. First prediction of sequence-specific recognition of double-helix nucleic acids by proteins. *Nat. Rev. Mol. Cell Biol.* 23:166.
17. McClarin, J. A., C. A. Frederick, …, J. M. Rosenberg. 1986. Structure of the DNA-Eco RI endonuclease recognition complex at 3 Å resolution. *Science.* 234:1526–1541.
18. Kim, Y. C., J. C. Grable, …, J. M. Rosenberg. 1990. Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing. *Science.* 249:1307–1309.
19. Wolberger, C., A. K. Vershon, …, C. O. Pabo. 1991. Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell.* 67:517–528.
20. Rich, A. 1992. Molecular recognition between proteins and nucleic acids. *In* The Chemical Bond: Structure and Dynamics. A. Zewail, ed. Academic Press Inc., pp. 31–86.
21. Otwinowski, Z., R. W. Schevitz, …, P. B. Sigler. 1988. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature.* 335:321–329.
22. Luscombe, N. M., S. E. Austin, …, J. M. Thornton. 2000. An overview of the structures of protein-DNA complexes. *Genome Biol.* 1. reviews001.1.
23. Garvie, C. W., and C. Wolberger. 2001. Recognition of specific DNA sequences. *Mol. Cell.* 8:937–946.
24. Wolberger, C. 2021. How structural biology transformed studies of transcriptional regulation. *J. Biol. Chem.* 296:100741.
25. Luisi, B. F., W. X. Xu, …, P. B. Sigler. 1991. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature.* 352:497–505.
26. Pavletich, N. P., and C. O. Pabo. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science.* 252:809–817.
27. Ellenberger, T. E., C. J. Brandl, …, S. C. Harrison. 1992. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell.* 71:1223–1237.
28. Raumann, B. E., M. A. Rould, …, R. T. Sauer. 1994. DNA recognition by β-sheets in the Arc repressor-operator crystal structure. *Nature.* 367:754–757.
29. Muller, C. W., F. A. Rey, …, S. C. Harrison. 1995. Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature.* 373:311–317.
30. Newman, M., T. Strzelecka, …, A. K. Aggarwal. 1995. Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science.* 269:656–663.
31. Luscombe, N. M., R. A. Laskowski, and J. M. Thornton. 2001. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* 29:2860–2874.
32. Jakubec, D., R. A. Laskowski, and J. Vondrasek. 2016. Sequence-specific recognition of DNA by proteins: binding motifs discovered using a novel statistical/computational analysis. *PLoS One.* 11:e0158704.
33. Kim, Y., J. H. Geiger, …, P. B. Sigler. 1993. Crystal structure of a yeast TBP/TATA-box complex. *Nature.* 365:512–520.
34. Kim, J. L., D. B. Nikolov, and S. K. Burley. 1993. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature.* 365:520–527.
35. Guzikevich-Guerstein, G., and Z. Shakked. 1996. A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nat. Struct. Biol.* 3:32–37.
36. Davey, C. A., D. F. Sargent, and T. J. Richmond. 2002. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* 319:1097–1113.
37. Lamoureux, J. S., J. T. Maynes, and J. N. Glover. 2004. Recognition of 5′-YpG-3′ sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. *J. Mol. Biol.* 335:399–408.
38. Schultz, S. C., G. C. Shields, and T. A. Steitz. 1991. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science.* 253:1001–1007.
39. Patra, A., L. D. Nagy, …, M. Egli. 2014. Kinetics, structure, and mechanism of 8-oxo-7,8-dihydro-2′-deoxyguanosine bypass by human DNA polymerase eta. *J. Biol. Chem.* 289:16867–16882.
40. Vasilyev, N., A. Polonskaia, …, A. Serganov. 2015. Crystal structure reveals specific recognition of a G-quadruplex RNA by a beta-turn in the RGG motif of FMRP. *Proc. Natl. Acad. Sci. USA.* 112:E5391–E5400.
41. Swan, M. K., R. E. Johnson, …, A. K. Aggarwal. 2009. Structure of the human Rev1-DNA-dNTP ternary complex. *J. Mol. Biol.* 390:699–709.
42. Hoffman, M. M., M. A. Khrapov, …, A. D. Ellington. 2004. AANT: the amino acid-nucleotide interaction database. *Nucleic Acids Res.* 32:D174–D181.
43. Park, B., H. Kim, and K. Han. 2014. DBBP: database of binding pairs in protein-nucleic acid interactions. *BMC Bioinf.* 15:S5.
44. Lustig, B., S. Arora, and R. L. Jernigan. 1997. RNA base-amino acid interaction strengths derived from structures and sequences. *Nucleic Acids Res.* 25:2562–2565.
45. Burke, B., A. Songon, and K. Musier-Forsyth. 2008. Functional guanine-arginine interaction between $tRNA^{Pro}$ and prolyl-tRNA synthetase that couples binding and catalysis. *Biochim. Biophys. Acta.* 1784:1222–1225.
46. Shashank Chavali, S., C. E. Cavender, …, J. E. Wedekind. 2020. Arginine forks are a widespread motif to recognize phosphate backbones and guanine nucleobases in the RNA major groove. *J. Am. Chem. Soc.* 142:19835–19839.
47. Klein, D. J., T. M. Schmeing, …, T. A. Steitz. 2001. The kink-turn: a new RNA secondary structure motif. *EMBO J.* 20:4214–4221.
48. Knight, R. D., and L. F. Landweber. 1998. Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem. Biol.* 5:R215–R220.
49. Reiss, C. W., Y. Xiong, and S. Strobel. 2017. Structural basis for ligand binding to the guanidine-I riboswitch. *Structure.* 25:195–202.
50. Reiss, C. W., and S. Strobel. 2017. Structural basis for ligand binding to the guanidine-II riboswitch. *RNA.* 23:1338–1343.
51. DeRouchey, J., B. Hoover, and D. C. Rau. 2013. A comparison of DNA compaction by arginine and lysine peptides: a physical basis for arginine rich protamines. *Biochemistry.* 52:3000–3009.
52. Wilds, C. J., M. A. Maier, …, M. Egli. 2002. Direct observation of a cytosine analogue that forms five hydrogen bonds to guanosine: guanidino G-clamp. *Angew. Chem. Int. Ed.* 41:115–117.
53. Wilds, C. J., M. A. Maier, …, M. Egli. 2003. Structural basis for recognition of guanosine by a synthetic tricyclic cytosine analogue: guanidinium G-clamp. *Helv. Chim. Acta.* 86:966–978.
54. Aggarwal, A. K., D. W. Rodgers, …, S. C. Harrison. 1988. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science.* 242:899–907.
55. Namba, K., R. Pattanayek, and G. Stubbs. 1989. Visualization of protein-nucleic acid interactions in a virus. Refined structure of intact

tobacco mosaic virus at 2.9 A resolution by X-ray fiber diffraction. *J. Mol. Biol.* 208:307–325.

56. Akhtar, A. 2021. The singularity is here. *The Atlantic.* 328:17–21.

57. Seeman, N. C. 1982. Nucleic acid junctions and lattices. *J. Theor. Biol.* 99:237–247.

58. Nguyen, N., J. J. Birktoft, and N. C. Seeman. 2012. The absence of tertiary interactions in a self-assembled DNA crystal structure. *J. Mol. Recogn.* 25:234–237.

59. Zheng, J., J. J. Birktoft, …, N. C. Seeman. 2009. From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature.* 461:74–77.