PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

Interaction Preferences between Nucleobase Mimetics and Amino Acids in Aqueous Solutions

Matea Hajnic, Juan I. Osorio[#] and Bojan Zagrovic[*]

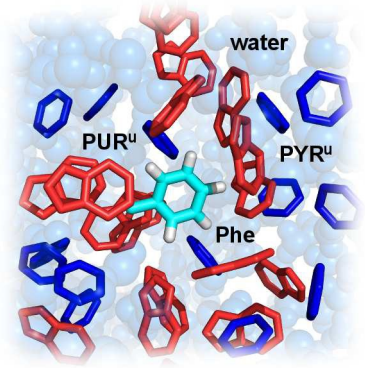
Department of Structural and Computational Biology, Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter 5, Vienna 1030, Austria

June 12th 2015

[#]present address: Juan I. Osorio, Institute for Theoretical Physics, ETH Zürich 8093, Switzerland

[*]to whom correspondence should be addressed. Tel: +43 1 4277 52271; Fax: +43 1 4277 9522; email: bojan.zagrovic@univie.ac.at

TABLE OF CONTENTS ENTRY



Interaction free energies between amino acids and nucleobase mimetics (unsubstitued purine or pyrimidine rings) derived from MD simulations reveal the influence of ring architecture on the specificity in amino-acid/RNA-nucleobase interactions.

ABSTRACT

Despite the paramount importance of protein-nucleic acid interactions in different cellular processes, our understanding of such interactions at the atomistic level remains incomplete. We have used molecular dynamics (MD) simulations and 15 μs of sampling time to study the behavior of amino acids and amino-acid sidechain analogs in aqueous solutions of different mimetics of naturally occurring nucleobases, including dimethylpyridine (DMP) and unsubstituted purine and pyrimidine rings. By using structural and energetic analysis, we have derived preference scales for the interaction of amino acids and their sidechain analogs with different nucleobase mimetics and have exhaustively compared them with each other. A close correspondence with a standard hydrophobicity measure in the case of the pyrimidine mimetic DMP and purines suggests that the hydrophobic effect is the main defining factor behind such interactions. We analyze our findings in the context of the origin of the genetic code and the recently proposed cognate mRNA-protein complementarity hypothesis. Most importantly, we show that unsubstituted purine and pyrimidine rings alone cannot differentiate between predominantly purine- and pyrimidine-coded amino acids, suggesting that for such specificity to exist, it must primarily reside in ring substituents.

INTRODUCTION

Interaction preferences between individual amino-acid residues and DNA or RNA nucleobases represent an important, reductionist foundation for understanding DNA-protein or RNA-protein interactions in general. Such preferences are an essential element in defining the specificity of recognition events between nucleic acids and proteins, and it is, therefore, important to fully understand their physico-chemical foundation at the atomistic level. From *ab initio* quantum-mechanical calculations[1–4] to statistical analyses of known nucleic-acid/protein complexes[5–14], significant progress in this direction has been made using computational approaches, with experimental studies lagging somewhat behind[15]. In particular, to the best of our knowledge, there exist no complete, experimentally derived scales capturing the affinity of individual amino acids for different biologically relevant nucleobases. The examples that are available invariably concern limited subsets of bases and amino acids only[15], and the reason can primarily be traced back to the solubility issues associated with naturally occurring nucleobases[16].

For this same reason, however, there has been a long tradition among experimentalists of studying the interactions between amino acids and different nucleobase mimetics. Carl Woese and coworkers, in particular, have championed the usage of substituted pyridines such as dimethylpyridine (DMP) for this purpose[17,18]. In a series of studies going back to 1960s and 1970s, they analyzed the chromatographic mobility of amino acids in aqueous solutions of pyridines and established the so-called "polar requirement" (PR) scale capturing the relative strength of interaction between the two types of molecules[17,18]. More recently, Mathew et al. have used molecular dynamics (MD) simulations to derive an updated version of the original experimental PR scale, with a relevant change seen only in the case of tyrosine (Pearson R between the two scale was R = 0.95)[19]. Importantly, by mapping their PR scale onto the genetic

code, Woese and coworkers made a far-reaching observation that amino acids with a similar value of PR are encoded by similar codons and vice versa[17]. They interpreted this as evidence for the stereochemical foundation of the genetic code, the idea that the code evolved as a consequence of direct interactions between amino acids and different nucleobases[17,18,20,21].

Recently, we have put this largely qualitative observation on a more quantitative footing and have shown that indeed there is a statistically significant level of correlation between the pyrimidine content of individual codons and the PR of their cognate amino acids[22]. More importantly, we have shown that pyrimidine density profiles of naturally occurring mRNA coding sequences closely mirror the PR profiles of their cognate proteins, with the absolute value of the median Pearson correlation coefficient in the case of the human proteome of 0.74[22]. We have used this to generalize the stereochemical hypothesis to the level of complete mRNA and protein sequences and suggest that the two may be physico-chemically complementary to each other and bind[22,23]. An important assumption in this claim was that the PR scale, which was derived using DMP, is a quality proxy for the scale capturing the interaction preferences of amino acids with standard RNA/DNA pyrimidines. This assumption has more recently been tested directly[14]. By analyzing a large set of known structures of RNA-protein complexes and deriving knowledge-based interaction preference scales for amino acids and different RNA nucleobases, we have indeed found support for the complementarity idea in the case of pyrimidines, but also extended it to the case of purines[14]. Specifically, we have shown that the profiles capturing the affinity of amino acids for guanines on the side of proteins closely match purine density on the side of their cognate mRNAs[14]. Interestingly, the opposite behavior was demonstrated in the case of adenines[14].

An important open question in this context concerns the physico-chemical foundation of individual base/amino-acid interaction preferences. In particular, what is the concrete influence of the nitrogenous-base ring architecture and ring substituents on these preferences? Do unsubstituted purine and pyrimidine rings alone already exhibit differentiated specificity in their interactions with amino acids? How does the picture change upon the addition of ring substituents? In order to address these questions, we employ here molecular dynamics (MD) simulations and extensive sampling to study the behavior of amino acids and amino acids sidechain analogs in aqueous solutions of DMP, unsubstituted purine (PUR^u) rings, unsubstituted pyrimidine (PYR^u) rings or a mixture of PUR^u and PYR^u rings (Figure 1), all at physically achievable concentrations. We use GROMOS 53A6 force field[24] for our simulations since this classical, united-atom force field was parameterized to accurately capture the hydrophobicity of amino-acid sidechain analogs, a property which is known to play an important role in amino-acid/nucleobase interactions. Specifically, we perform structural and energetic analysis on the simulated systems and derive scales of interaction preferences between the participating groups. We test and validate our simulation methodology by comparing the simulation-based DMP scales with Woese's PR scale[18], the only available experimental scale capturing the affinity of all natural amino acids for base-like compounds. The derived DMP, PUR^u and PYR^u scales allow us to study the influence of the nitrogenous-base ring architecture and substituents on base/amino-acid interactions and study their physico-chemical basis. In particular, we compare and contrast the derived scales with the results of a related analysis performed using naturally occurring RNA nucleobases[25]. Finally, we critically discuss the cognate mRNA/protein complementarity hypothesis[14,22,23] in light of the derived scales.

RESULTS

The experimental PR scale[17,18], derived by analyzing the chromatographic mobility of amino acids in water mixtures of substituted pyridines, is one of the few examples where interactions between amino acids and nitrogenous bases have been systematically explored in experiment. In order to compare and validate our methodology and results against experimental ones, we have first simulated individual amino acids and their sidechain analogs in the aqueous solution of DMP (Figure 2A), one of the derivatives used in the experiments by Woese and coworkers[18] (Figure 1). In the experiment, the PR of a given amino acid was defined as the slope of the linear fit between the logarithm of its retention coefficient R and the logarithm of the mole fraction of water in the pyridine-water solvent. Specifically, amino acids with low PR (Ile, Leu, Phe, Met, Trp, Cys) exhibit a higher propensity to interact with DMP than the ones with high PR (Asp, Glu, Lys, Asn, Arg, Gln)[18]. We have calculated radial distribution functions (RDFs) from our simulations to analyze the preferences of different amino acids to interact with DMP (Figure S1 in the Supporting Information). Overall, our analysis demonstrates that amino acids with low experimentally measured PR (Phe, Leu, Trp) indeed exhibit a higher probability to be found in the close vicinity of DMP molecules than those with high PR (Glu, Asp, Lys, Asn, Gln, Arg) (Figure S1). The same behavior is observed for their sidechain analogs as well (Figure S3). In order to illustrate this finding, in Figure 2B we show the RDFs of Phe and Glu, two amino acids whose probabilities of being preferentially surrounded by DMP as compared to water molecules are, respectively, high and low (RDFs for other amino acids are given in Figure S1). While the first peak in the Phe/DMP RDFs is centered at approximately 0.6 nm, in the case of Glu/water RDFs it is at approximately 0.4 nm. This difference is a consequence of the smaller size of Glu on the one hand and the specific structural arrangement of closely interacting Phe and DMP pairs

on the other. The latter predominately exhibit a stacked geometry (27.6%) with a smaller fraction adopting a "T-shape" geometry (6.4%) as calculated using a 0.6 nm distance cutoff for the separation between the Phe and DMP centers of mass (Figure S2). When it comes to the Phe sidechain analog, its probability to be found close to DMP molecules is even higher than for the complete amino acid (Figure 2C). This is a consequence of the fact that in our simulations amino acids have charged carboxylic and amino termini, which prevents their hydrophobic sidechains from further entering into the more hydrophobic DMP phase in the simulated systems. A similar situation is seen for a number of other sidechain analogs as well (Figure S3).

To put the RDF probabilities on a quantitative footing, for each amino acid and sidechain analog we have calculated its interaction free energy with DMP relative to that with water (Eq. 2 and 3) as well as a difference between the total potential energy of interaction with DMP and water (Eq. 4). In Figure 3 we plot the thus-obtained $\Delta\Delta G_{NB-W}$ for amino acids against the experimentally derived PR scale ($PR_{experiment}$)[18] and the two exhibit good agreement and a strong linear relationship with a Pearson correlation coefficient of R = 0.86 (individual scales given above the graph). Although the two scales were derived in very different ways, the good agreement between them is a testament to the quality of our force field and the general computational methodology used. The equivalent enthalpy-based scale (ΔE_{aa}) agrees with the experimental scale even better (R = 0.93, Figure 3 inset). Similarly, the scales obtained for sidechain analogs also agree closely with the experimental PR scale (R=0.85 for the free energy scale and R=0.96 for the enthalpy scale) which is not surprising given that the computational scales for amino acids and sidechain analogs correlate closely with each other (R = 0.81 for the free energy scale and R=0.97 for the enthalpy scale). Moreover, both amino-acid scales agree closely with an analogous computational scale of amino-acid binding propensities for DMP, derived by Mathew

et al.[19] using an RDF-based formalism and the CHARMM 27 force field[26] (Pearson R = 0.82 for the free energy scale and R = 0.93 for the enthalpy scale). As sidechain analogs model the behavior of protein residues at RNA-protein interfaces arguably better as compared to the zwitterionic amino acids, in the rest of the text we focus on sidechain analogs, while always also providing the results for amino acids for comparison.

The original PR scale was derived using pyridines and not the biologically more appropriate pyrimidines. In order to test how interchangeable the two chemicals indeed are as well as to further probe the dependence of the derived scales on the specific structure of the heterocyclic ring, we have repeated the above analysis for amino acids and their sidechain analogs in aqueous solutions of unsubstituted pyrimidine (PYR^u) or purine (PUR^u) rings (Figure 1). Remarkably, the obtained relative interaction free energy scales for PUR^u correlate linearly and directly with the DMP scales with the Pearson correlation coefficient of R = 0.84 for complete amino acids and R = 0.85 for sidechain analogs (see Figure S4 and Table S1 for details in the Supporting Information). The equivalent correlations are much weaker for PYR^u (R = 0.36 for complete amino acids and R = 0.13 for sidechain analogs), whereas the main difference arises primarily from the low dynamic range of amino acids' affinities for PYR^u, with Trp affinity being only exception (Figure S4, Table S1). Furthermore, the RDFs show that indeed sidechain analogs whose corresponding amino acids exhibit a high preference for DMP (e.g. Trp) also prefer unsubstituted purines over water and *vice versa* (Figure S5). In general, there appears to be little qualitative difference in the behavior of amino acids or sidechain analogs between DMP and PUR^u solutions, suggesting that the chemical differences between these bases do not significantly affect the way they interact with amino acids or their sidechains in crowded solutions. On the contrary, PYR^u solutions exhibit a qualitatively significantly different behavior in this regard,

which can be traced back to their microscopic structure. Namely, while DMP and PUR^u aqueous solutions exhibit a largely biphasic behavior, with bases grouping together, PYR^u molecules tend to be individually highly solvated and do not closely associate with other bases. More specifically, the base-base RDFs in DMP and PUR^u solutions exhibit multiple well-defined peaks, with the first peak at a short distance of < 1 nm in both cases, indicating close interactions (Figure S6 A and C). However, there appear to be no significant peaks in DMP- or PUR^u-water RDFs (Figure S6 A and C). In PYR^u solutions, on the other hand, both base-base and base-water RDFs have a well-defined first peak within 1 nm, indicating the presence of highly hydrated clusters (Figure S6 B). Moreover, the base-water RDFs in the DMP and PUR^u systems increase linearly with distance, consistent with the existence of a well-defined boundary between two phases (Figure S6 A and C). As a consequence, the preferences of amino acids and their side chain analogs for DMP or PUR^u are primarily defined by the solvent entropy. By introducing amino acids/side-chain analogs, one disrupts a dynamic hydrogen bond network of the surrounding water molecules. In order to minimize the disruption, water molecules organize themselves around more hydrophobic domains in cage-like structures to preserve the hydrogen-bonded network at the expense of translational and rotational entropy. In nitrogenous base-water mixtures, amino acids interact with bases present in the solution and in this way minimize their exposed hydrophobic area to the surroundings and reduce the total number of water molecules engaged in the formation of a water cage around them. The given effect is stronger for the more hydrophobic PUR^u that forms bigger hydrophobic patches in the solvent, than for the less hydrophobic PYR^u [27].

How pronounced are the relative preferences of different sidechain analogs to interact with pyrimidines or purines? In order to address this question, we have examined the behavior of

sidechain analogs in simulated water mixtures of PUR^u and PYR^u rings. Already from a qualitative analysis of such mixed systems, it is apparent that some sidechain analogs tend to interact more with PUR^u and some with PYR^u. In Figures 4A and 4B, we show representative snapshots and the associated RDFs from the simulations of two sidechain analogs that exhibit a notable difference in their relative free energies of interaction with a nitrogenous base: Phe, which is mostly surrounded by PUR^u, and Lys, which mostly resides in the vicinity of PYR^u. Moreover, we calculate contact coefficients (CCs) for every sidechain analog by enumerating the number of interatomic contacts between a sidechain analog and PYR^u or PUR^u molecules in the desired distance range (0.35 nm). If CC > 1, a given sidechain analog interacts more with PUR^u than with PYR^u and *vice versa* (Figure 4C). All charged and polar sidechain analogs (Lys, Asp, Glu, Arg, Ser, Thr, Asn and Gln) display more interatomic contacts with PYR^u, while the highest propensity for interacting with PUR^u is exhibited by the aromatic Phe and Trp (Figure 4C). In general, the free energy-based sidechain analog interaction propensities for PUR^u vs. PYR^u derived from mixed systems show a very similar ranking of preferences as CCs (Spearman rank coefficient $\rho = -0.90$ for PYR^u and $\rho = -0.99$ for PUR^u), suggesting that these two measures are mutually consistent for mixed systems (Table S2).

Despite the differences in their microscopic origin, the relative free energy scales for interaction with PUR^u and PYR^u derived from individual systems correlate moderately with each other with a Pearson correlation coefficient R = 0.58 for amino acids or R = 0.46 for sidechain analogs (Table S1). This similarity is even more pronounced in the case of the PUR^u/PYR^u mixture with R = 0.76 for amino acids or R = 0.70 for sidechain analogs (Figure 4D, Table S1). Finally, the PUR^u or PYR^u sidechain analog interaction free energy scales derived from mixed systems correlate closely with the equivalent scales derived from two separated systems where only one

base type is present (R = 0.97 for PUR^u and R = 0.98 for PYR^u, Figure 4E; for amino acid scales R = 0.87 and R = 0.94, respectively, Table S1). This close correspondence is remarkable and it suggests that relative free energy scales for the interaction between sidechain analogs or amino acids and different nitrogenous bases may in general be obtained from individual scales derived from simulations or experiments with single nitrogenous base types only.

What is the physicochemical nature of the interactions studied herein? To address this question, we have compared the derived scales with five scales obtained by Atchley and coworkers using a multivariate factor analysis of approximately 500 different amino-acid property scales[28] (Figure 5A). These 5 scales are largely independent from each other and capture a range of important amino-acid properties including: hydrophobicity (Factor I scale), secondary structure propensity (Factor II scale), molecular volume (Factor III scale), codon diversity (Factor IV scale) and electrostatic charge (Factor V scale). Notably, the derived interaction free energy sidechain analog scales for DMP and PUR^u exhibit a strong correlation with the Factor I hydrophobicity scale, with Spearman ρ coefficients of 0.86 and 0.81 for DMP and PUR^u scales, respectively (Figure 5A), and little or no correlation with any other factor scales ($|\rho| \leq 0.4$ in all cases, Figure 5A). The sidechain PYR^u scale exhibit no significant correlation with the Factor I scale (ρ = 0.22), but does correlate somewhat with the Factor IV scale (ρ = 0.53). As for the outliers, Trp appears to be the only consistent one when comparing the PYR^u propensity scale with the five Factor scales, but this is the only commonality that we could detect. Expectedly, the above analysis is consistent with results obtained when comparing the DMP, PUR^u and PYR^u scales with the entire collection of over 500 amino-acid scales (Figure 5A inset). Namely, hydrophobicity-related amino-acid scales (152 in total) are strongly correlated with DMP and PUR^u scales on average (median $\rho^2 > 0.5$, Figure 5A inset), whereas all the remaining amino-acid

scales which are not related to hydrophobicity, do not exhibit any significant levels of correlation on average (388 scales in total, median $\rho^2 \leq 0.17$). Conversely, the PYR^u scales exhibit a pronouncedly idiosyncratic behavior with no other known amino-acid scales correlating with them significantly.

DISCUSSION

Our results show that different amino acids and their sidechain analogs exhibit clearly defined interaction preferences for nucleobase mimetics and that, in the case of DMP or PUR^u, these preferences are closely related to amino acid hydrophobicity. Namely, hydrophobic amino acids tend to interact more closely with DMP or PUR^u, regardless of their specific nature, while hydrophilic amino acids prefer to interact with water. Our results also show that amino acids and their sidechain analogs display highly differentiated interaction propensities for different nucleobase mimetics depending on the ring architecture. The most direct evidence for this is provided by our MD simulations of mixed systems, where we could directly measure relative interaction free energies between amino acids or their sidechain analogs and PYR^u and PUR^u present in the mixture at the same time. While the classical force field employed in the present study by definition cannot fully account for all of the physics of amino-acid/nitrogenous base interactions, a strong agreement with experimental results in the case of DMP suggests that the general trends in interaction specificity are well captured using such a model. Furthermore, various studies have used GROMOS 53A6 force field to examine transitions between conformation states of a protein upon ligand binding[29] and intermolecular interactions responsible for stability of protein complexes[30] or to reconcile initially inconsistent experimental results[31]. Importantly, in all of these examples, a satisfactory agreement has been found between experimental results and simulations in multiple respects. Additionally, in a recent study, we have

determined the absolute binding free energies between RNA/DNA nucleobases and amino-acid side-chain analogs using the same force field as here[25], and have demonstrated a close correspondence with the available experimental data[15] when it comes to relative ranking. For example, the largest available experimental set included the GUA affinities of 8 selected amino acids[15] and there we observed a correlation with the simulations performed using the GROMOS 53A6 nucleobase set with a Pearson R=0.87[25].

Carl Woese and co-workers have used amino-acid PR measurements to hypothesize that the universal genetic code may have originated from direct interactions between amino acids and their cognate codons («the stereochemical hypothesis»)[17,18,20,21]. Moreover, we have recently expanded this proposal and provided evidence that proteins and their cognate mRNAs may in general be physico-chemically complementary to each other and bind, especially if unstructured[14,22,23]. An important assumption behind both Woese's and our argumentation is that DMP is a representative proxy for naturally occurring pyrimidines, at least when it comes to interactions with amino acids or their sidechain analogs. Our present findings provide important new information concerning this assumption. In particular, our results suggest that there is little difference in the behavior of amino acids or their sidechain analogs when it comes to their interactions with DMP on the one hand and, surprisingly, PUR^u on the other. In fact, non-polar sidechain analogs, encoded by pyrimidine-rich codons, exhibit higher interaction propensities for PUR^u than for PYR^u, going against both Woese's assumptions and our interpretation[14,22,23] of the experiments performed by him and coworkers[17,18]. On the other hand, knowledge-based interaction preference scales of amino acids and RNA nucleobases[14] derived recently by us as well as umbrella sampling MD calculations[25] and direct simulations of amino acids and their sidechain analogs in water solutions of RNA nucleobases[32], all consistently suggest that the key

element in determining the specificity of interaction between amino acids and bases is not the nature of the heterocyclic ring, but rather that of ring substituents. In agreement with this, when comparing the affinity scales derived in this study with amino-acid sidechain analog affinities for realistic RNA nucleobases[25] (Table S3), PUR^u affinity scales correlate strongly with all four amino-acid RNA affinity scales, while PYR^u affinity scales do not correlate closely with any of the RNA affinity scales. What is more, we have recently shown that several sidechain analogs exhibit opposite behavior in GUA- and ADE-based solutions with, for example, Lys and Arg being among the strongest interacting partners of GUA and among the weakest interacting partners of ADE[32]. In Figure 5B, we compare the relative preferences of Lys for PUR^u and PYR^u derived presently with its relative interaction preferences for GUA/CYT, GUA/URA, ADE/CYT or ADE/URA derived from the corresponding absolute binding free energies obtained recently using umbrella-sampling simulations[25]. While the ADE-URA relative scale shows similar qualitative behavior to the relative PUR^u-PYR^u scale, suggesting that in this case interactions with nucleobase rings appear to be more important, the GUA/CYT, GUA/URA and ADE/CYT relative values exhibits the opposite behavior (Figure 5B), reconfirming the importance of ring substituents in defining nucleobase-amino acid interactions.

In addition to shedding further light on the mRNA-protein complementarity hypothesis[14,22,23] as discussed above, we believe our present results also provide an important advance in a more methodological context. Namely, nucleobase-like compounds are traditionally not a typical choice for deriving amino-acid hydrophobicity scales in experiment, with ethanol, octanol or cyclohexane being more frequently used[33–35]. However, a close correspondence between the Factor I hydrophobicity scale[28] and the PUR^u scale derived herein demonstrates that latter is likely to a large extent based precisely on hydrophobic interactions. Given the paramount

biological significance of amino acid-nucleobase i.e. protein-nucleic acids interactions, we would like to suggest that nucleobases and nucleobase-like compounds, and especially unsubstituted purine rings, can actually be seen as natural and highly relevant non-polar compounds to be used in the determination of amino-acid hydrophobicity scales. However, the more complex behavior of real nucleobases such as guanine or adenine, together with the solubility issues associated with them, suggest that these may be less well-suited for this purpose.

METHODS

Molecular dynamics simulations

We have used MD simulations to study the behavior of the 20 natural amino acids and 18 of their sidechain analogs (all except for Gly and Pro) in aqueous solutions of either 2,6-dimethylpyridine (DMP), unsubstituted pyrimidine (PYR^u) or unsubstituted purine (PUR^u) (Figure 1). Moreover, we have also studied amino acids and their sidechain analogs in a mixture of PUR^u and PYR^u. All simulations were carried out using the Gromacs 4.5.1. simulation package[36], united-atom GROMOS 53A6 force field[24] and SPC/E water model[37]. In all simulations, a single amino acid or a sidechain analog (corresponding to an amino-acid residue with a backbone part of the amino acid replaced by a hydrogen atom) was placed in the center of a cubic box (initial size 4x4x4 nm^3) with nitrogenous bases and water molecules distributed around it in random orientations so as to achieve the molar fraction of water of 0.86. In total, there were approximately 1000 molecules in each system: one amino acid or sidechain analog, 140 nitrogenous base molecules and the rest water molecules (Table S4). In a mixed system with both PUR^u and PYR^u molecules present, amino acids or sidechain analogs were surrounded by 859 water, 61 PUR^u and 79 PYR^u molecules in a cubic box (initial size 4x4x4 nm^3) resulting again in the molar fractions of water of 0.86. The PUR^u/PYR^u ratio in these simulations was chosen so as to obtain a similar total

number of atoms for the two nitrogenous bases. Importantly, in all three cases, our simulations were carried out at physically achievable concentrations of bases. All amino acids were simulated in their zwitterionic form corresponding to a pH of 7. In the case of charged amino acids or sidechain analogs, a randomly selected water molecule was replaced by a counter ion (Na^+ or Cl^-) in order to achieve electrical neutrality.

All simulations were carried out using a 2 fs integration step in the case of DMP systems or a 1 fs integration step for PUR^u and/or PYR^u systems. The parameters for nitrogenous bases were derived from those corresponding to the most similar nucleotides in the GROMOS 53A6 force field. As the GROMOS force field does not have a standard set of parameters for DMP, appropriate force field parameters were derived from the automated topology builder (ATB)[38,39]. The partial charges assigned by ATB differ slightly from parameters, which would be assigned according to analogy to existing RNA nucleobases in GROMOS. In particular, a smaller partial charge on nitrogen atom in the DMP molecule would be expected in the latter case. However, for aromatic systems, the general transfer-rules from one molecule to another one are weak in GROMOS because of the great difference in charge distribution over the ring. In our case, pyrimidine molecule has an additional nitrogen atom in the aromatic ring as compared to DMP and this can significantly change the charge distribution in the entire ring and make parameters not transferable. This is the main reason why for DMP we have opted to use ATB parameterization. All bonds were constrained using LINCS[40]. Long-range electrostatic interactions were treated using Particle Mesh Ewald (PME) summation with grid spacing of 0.12 nm in the case of DMP systems and 0.14 nm for PUR^u and/or PYR^u systems with interpolation order of 4, while cut-offs for short-range Coulombic and van der Waals interactions were set to 0.9 nm. The temperature and pressure were kept at 300 K and 1 bar throughout, using V-rescale

thermostat (τ_T = 0.1 ps)[41] and Parrinello-Rahman barostat (τ_p = 2 ps and compressibility = 4.5 x 10^{-5} bar^{-1})[42], respectively. Prior to running all simulations, we have tested two different compressibility values, one reported here, and another one that was calculated by weighting the compressibility values of water and DMP by their molar mass ratio in the system. However, no matter what compressibility value was used, there was no significant difference in the thermodynamic properties of the simulated system. After minimization using the steepest descent algorithm in water (10000 – 25000 steps), the systems with PUR^u and/or PYR^u were equilibrated in the NPT ensemble for 1.2 ns with position restraints placed on the amino acid or sidechain analog. In addition to the same minimization procedure, the systems with DMP were first equilibrated in the NVT ensemble for 800 ps and then subjected to 400 ps of equilibration in the NPT ensemble with the same position restraints placed on the amino acid or sidechain analog. All production runs, each 100 ns long, were performed in the NPT ensemble for a total of 15.2 μs of simulated time over all systems (20 amino acids x 4 system setups x 100 ns + 18 sidechain analogs x 4 system setups x 100 ns = 15.2 μs). The simulated trajectories were analyzed both structurally and energetically. For simplicity, we describe the procedure for sidechain analogs only, but equivalent calculations were also carried out for all amino acid-containing systems.

Structural analysis of trajectories

For structural analysis, radial distribution functions were calculated by using as reference points the centers of mass of sidechain analogs, nitrogenous bases and water molecules. Furthermore, the formalism proposed by Stumpe and Grubmüller[43] was used to calculate contact coefficients (CC), a structural measure which quantifies the number of interatomic contacts between a sidechain analog and nitrogenous base ($N_{X\text{-}NB}$) and a sidechain analog and water ($N_{X\text{-}W}$) in a

desired contact range normalized by the total number of nitrogenous base (M_{NB}) and water (M_W) atoms:

$$CC_{NBW_x} = N_{X-NB} \cdot M_W / \left(N_{X-W} \cdot M_{NB} \right) \quad (1)$$

All of the reported results are given for the contact range of 0.35 nm. Multiple contact ranges were also explored (0.35 nm, 0.45 nm and 0.55 nm), but without observing any significant difference in the obtained contact coefficients.

Estimation of relative interaction free energies

In two- and three- component systems, preferential solvation can be treated using statistical-mechanical approaches based on Kirkwood-Buff (KB) integrals[44,45]. Briefly, preferential solvation of a solute (e.g. sidechain analogs) with respect to different components of the solvent (e.g. nitrogenous bases and water) can be derived by comparing the local solvent composition around the solute with the bulk solvent composition with the help of RDF-based KB integrals. Consequently, one can estimate the free energy of interaction between a sidechain analog X and different solution components (nitrogenous bases or water) by comparing the local concentration ($c^{local}_{solution-component}$) of solution components around the sidechain analog in question against their concentration in the bulk ($c^{bulk}_{solution-component}$):

$$\Delta G_{X,solution-component} = -k_B T \ln \frac{c^{local}_{solution-component}}{c^{bulk}_{solution-component}} \quad (2)$$

In order to estimate relative interaction free energies with nitrogenous bases and water for sidechain analogs, we have calculated the local concentration (c_{local}) of solvent components from the average number of solvent molecules of a given kind within the volume defined using a cut-off distance from the sidechain analog center of mass (r_{local}). Here, a solvent molecule is

considered to be within the local volume if any of its atoms can be found within the cutoff distance. The obtained sidechain analog free energies for nitrogenous bases and water were further subtracted from each other to obtain differences in free energy of sidechain analog-nitrogenous base and sidechain analog-water interactions:

$$\Delta\Delta G_{X,NB-W} = \Delta G_{X,NB} - \Delta G_{X,W} \quad (3)$$

Here, we report such relative free energies for r_{local} at which $\Delta G_{X,NB}$ reaches a minimum.

Estimation of interaction propensities

For the analysis of the enthalpic component of free energy, differences between the total force-field potential energies corresponding to sidechain analog-nitrogenous base ($E_{X,NB}$) and sidechain analog-water interactions ($E_{X,W}$) were calculated as time- and ensemble- averages:

$$\Delta E_{X,NB-W} = \left\langle E_{X,NB} - E_{X,W} \right\rangle_{time,ensemble} \quad (4)$$

To be able to compare systems with slightly different molar compositions, the calculated potential energies between sidechain analog and water molecules were rescaled before obtaining the interaction propensity scale so as to have all systems correspond to exactly 0.86 molar fraction water. The assumption behind such rescaling is that the few additional water molecules behave on average in the same way as the rest of the water molecules in the system and contribute to the overall sidechain analog-water potential energy proportionally to their number. Analogous structural and energetic analysis was performed for systems containing amino acids with contacts and interaction energies evaluated over all amino-acid atoms.

The obtained scales were compared to 540 different scales describing physico-chemical properties (of which 152 were hydrophobicity-related) of naturally occurring amino acids that

were derived from AAindex database[46] as previously described[22]. Amino-acid and sidechain analog relative interaction free energy scales, interaction propensity scales and contact coefficients are given in Table S5.

CONCLUSION

Motivated by the paramount importance of direct binding between nucleic acids and proteins in various cellular processes, we have here explored the basic physicochemical principles behind such interactions. Our main result shows that amino-acid preferences for nucleobases are strongly affected by the nature of the heterocyclic ring, but also by ring substituents, which may reverse the effects of ring architecture. These findings shed light on an important, foundational problem in biology, that of the origin of the genetic code, and provide physico-chemical constraints that any putatively stereochemical mechanism of code's development should obey. We hope that the scales derived herein will provide a rigorous basis for understanding and manipulating protein-nucleic acid interactions in both fundamental and applied contexts.

ACKNOWLEDGEMENTS

We thank members of the Laboratory of Computational Biophysics at MFPL for useful advice and critical reading of the manuscript. The funding by the European Research Council (Starting Independent grant 279408 to BZ) is also gratefully acknowledged.

REFERENCES

1 C. Biot, E. Buisine, J. M. Kwasigroch, R. Wintjens and M. Rooman, *J. Biol. Chem.*, 2002, **277**, 40816–40822.
2 L. R. Rutledge, L. S. Campbell-Verduyn, K. C. Hunter and S. D. Wetmore, *J. Phys. Chem. B*, 2006, **110**, 19652–19663.

3 L. R. Rutledge, H. F. Durst and S. D. Wetmore, *Phys. Chem. Chem. Phys.*, 2008, **10**, 2801–2812.
4 A. Ebrahimi, M. Habibi-Khorassani, A. R. Gholipour and H. R. Masoodi, *Theor. Chem. Acc.*, 2009, **124**, 115–122.
5 N. M. Luscombe, R. A. Laskowski and J. M. Thornton, *Nucleic Acids Res.*, 2001, **29**, 2860–2874.
6 M. Treger and E. Westhof, *J. Mol. Recognit.*, 2001, **14**, 199–214.
7 E. Jeong, H. Kim, S. W. Lee and K. Han, *Mol. Cells*, 2003, **16**, 161–167.
8 M. M. Hoffman, M. A. Khrapov, J. C. Cox, J. C. Yao, L. N. Tong and A. D. Ellington, *Nucleic Acids Res.*, 2004, **32**, D174–D181.
9 J. E. Donald, W. W. Chen and E. I. Shakhnovich, *Nucleic Acids Res.*, 2007, **35**, 1039–1047.
10 M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag and R. B. Altman, *RNA*, 2009, **15**, 189–199.
11 L. Pérez-Cano, A. Solernou, C. Pons and J. Fernández-Recio, *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*, 2010, 293–301.
12 J. Kondo and E. Westhof, *Nucleic Acids Res.*, 2011, **39**, 8628–8637.
13 I. Tuszynska and J. M. Bujnicki, *BMC Bioinformatics*, 2011, **12**, 348.
14 A. A. Polyansky and B. Zagrovic, *Nucleic Acids Res.*, 2013, **41**, 8434–8443.
15 P. Thomas and S. Podder, *Febs Lett.*, 1978, **96**, 90–94.
16 Y. S. Dannenfelser RM, *Coll. Pharm. Univ. Ariz. Tucson AZ*, 1992.
17 C. Woese, *Proc. Natl. Acad. Sci. U. S. A.*, 1965, **54**, 1546–1552.
18 C. R. Woese, *Naturwissenschaften*, 1973, **60**, 447–459.
19 D. C. Mathew and Z. Luthey-Schulten, *J. Mol. Evol.*, 2008, **66**, 519–528.
20 C. Woese, *Proc. Natl. Acad. Sci. U. S. A.*, 1968, **59**, 110–117.
21 C. Woese, *J. Mol. Biol.*, 1969, **43**, 235–240.
22 M. Hlevnjak, A. A. Polyansky and B. Zagrovic, *Nucleic Acids Res.*, 2012, **40**, 8874–8882.
23 A. A. Polyansky, M. Hlevnjak and B. Zagrovic, *RNA Biol.*, 2013, **10**, 1248–1254.
24 C. Oostenbrink, A. Villa, A. E. Mark and W. F. van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.
25 A. de Ruiter and B. Zagrovic, *Nucleic Acids Res.*, 2014, **43**, 708–718.
26 A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.
27 Y. C. Martin, *J. Med. Chem. - J MED CHEM*, 1996, **39**, 1189–1190.
28 W. R. Atchley, J. Zhao, A. D. Fernandes and T. Drüke, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 6395–6400.
29 Y. Yuan, D. W. Bleile, X. Wen, D. A. R. Sanders, K. Itoh, H. Liu and B. M. Pinto, *J. Am. Chem. Soc.*, 2008, **130**, 3157–3168.
30 J. A. Lemkul and D. R. Bevan, *J. Phys. Chem. B*, 2010, **114**, 1652–1660.
31 P. R. Wilderman, M. B. Shah, T. Liu, S. Li, S. Hsu, A. G. Roberts, D. R. Goodlett, Q. Zhang, V. L. Woods, C. D. Stout and J. R. Halpert, *J. Biol. Chem.*, 2010, **285**, 38602–38611.
32 M. Hajnic, J. I. Osorio and B. Zagrovic, *Nucleic Acids Res.*, 2014, **42**, 12984–12994.
33 Y. Nozaki and C. Tanford, *J. Biol. Chem.*, 1971, **246**, 2211–2217.
34 A. Radzicka and R. Wolfenden, *Biochemistry (Mosc.)*, 1988, **27**, 1664–1670.
35 W. C. Wimley, T. P. Creamer and S. H. White, *Biochemistry (Mosc.)*, 1996, **35**, 5109–5124.

36 B. Hess, *Abstr. Pap. Am. Chem. Soc.*, 2009, **4**, 435–447.
37 H. Berendsen, J. Grigera and T. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
38 A. K. Malde, L. Zuo, M. Breeze, M. Stroet, D. Poger, P. C. Nair, C. Oostenbrink and A. E. Mark, *J. Chem. Theory Comput.*, 2011, **7**, 4026–4037.
39 S. Canzar, M. El-Kebir, R. Pool, K. Elbassioni, A. E. Mark, D. P. Geerke, L. Stougie and G. W. Klau, *J. Comput. Biol.*, 2013, **20**, 188–198.
40 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.
41 G. Bussi, D. Donadio and M. Parrinello, *Abstr. Pap. Am. Chem. Soc.*, 2007, **126**.
42 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
43 M. C. Stumpe and H. Grubmüller, *J. Am. Chem. Soc.*, 2007, **129**, 16126–16131.
44 A. Ben-Naim, *Pure Appl. Chem.*, 1990, **62**, 25–34.
45 S. Weerasinghe and P. E. Smith, *J. Chem. Phys.*, 2003, **118**, 5901–5910.
46 S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, *Nucleic Acids Res.*, 2008, **36**, D202–205.

FIGURE CAPTIONS

Figure 1. Chemical structure of nitrogenous bases used in MD simulations.

Figure 2. Radial distribution functions of amino acids and their sidechain analogs in DMP-water mixtures. (**A**) A typical snapshot from the simulation with a single amino acid (Phe) in the DMP-water mixture. (**B**) Radial distribution function g(r) showing a higher probability of finding DMP molecules close to amino acid phenylalanine (Phe_{aa}) than to water and *vice versa* for glutamate (Glu_{aa}). (**C**) Radial distribution function g(r) of the corresponding amino-acid sidechain analogs (Phe_{sca} and Glu_{sca}).

Figure 3. Comparison between experimentally and computationally derived DMP affinities. Correlation between experimentally derived polar requirement ($PR_{experiment}$) scale[18] and the amino-acid relative interaction free energy scale for DMP (kJ/mol) obtained by simulation. Inset: correlation between the enthalpy-based amino-acid interaction propensity scale for DMP (kJ/mol) with the PR scale. Top: experimentally derived polar requirement

scale (PR)[18]; amino acid ($\Delta\Delta G_{aa}$) and sidechain analog ($\Delta\Delta G_{sca}$) interaction free energy scales for DMP (kJ/mol) derived from MD simulations.

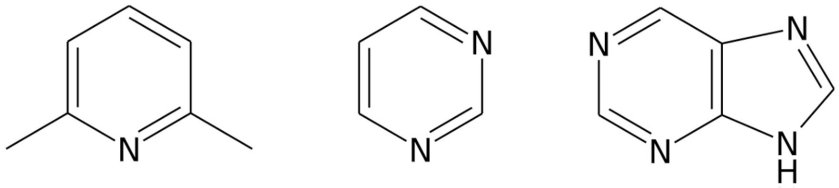
Figure 4. Propensities of amino-acid sidechain analogs for nucelobase mimetics PUR^u and PYR^u derived from mixed systems. (A) Snapshot of mixed systems with two different nitrogenous bases, PUR^u and PYR^u, present at the same time in the system. Phe is mostly surrounded by PUR^u, and Lys by PYR^u. (**B**) Radial distribution function g(r) showing higher probability of finding PUR^u close to Phe than PYR^u, and vice versa for system with Lys. (**C**) Contact coefficients for all 18 sidechain analogs derived from simulations of mixed systems with two different nitrogenous bases (PUR^u and PYR^u) present. Side chains with contact coefficients >1 prefer to interact with PUR^u, while side chains with contact coefficients <1 prefer to interact with PYR^u. (**D**) Correlation between the sidechain analog interaction free energy scale for PUR^u and the sidechain analog interaction free energy scale for PYR^u derived from mixed systems where both PUR^u and PYR^u are present at the same time. (**E**) Correlation between amino-acid sidechain analog relative interaction free energy scales for PUR^u and PYR^u derived from two separate systems ($\Delta\Delta G_{sca}$) and from the mixed system ($\Delta\Delta G^*_{sca}$) where both nitrogenous bases, PUR^u and PYR^u, were present at the same time.

Figure 5. Physicochemical nature of interactions between amino acids and nucleobase mimetics. (**A**) Spearman coefficients of correlations between 5 Factor scales[28] and amino-acid sidechain analog relative interaction free energy scales for DMP (yellow), PUR^u (red) and PYR^u (blue), derived from MD simulations. Inset: probability density distributions of ρ^2 for correlations between amino-acid sidechain analog relative interaction free energy scales for DMP (top), PYR^u (middle) and PUR^u (bottom), derived from MD simulations, and 152

hydrophobicity-related scales (solid line) or 388 non-hydrophobicity-related amino-acid property scales (dashed line). (**B**) Relative interaction preference of amino-acid sidechain analog Lys for PUR^u and PYR^u derived in this study (left) and relative interaction preferences of amino-acid sidechain analog Lys for standard nucleobases – GUA/CYT, GUA/URA, ADE/CYT or ADE/URA derived from the corresponding absolute binding free energies obtained using umbrella-sampling simulations[25].

FIGURES

Figure 1



2,6-DIMETHYLPYRIDINE (DMP) PYRIMIDINE (PYR[u]) PURINE (PUR[u])

Figure 2

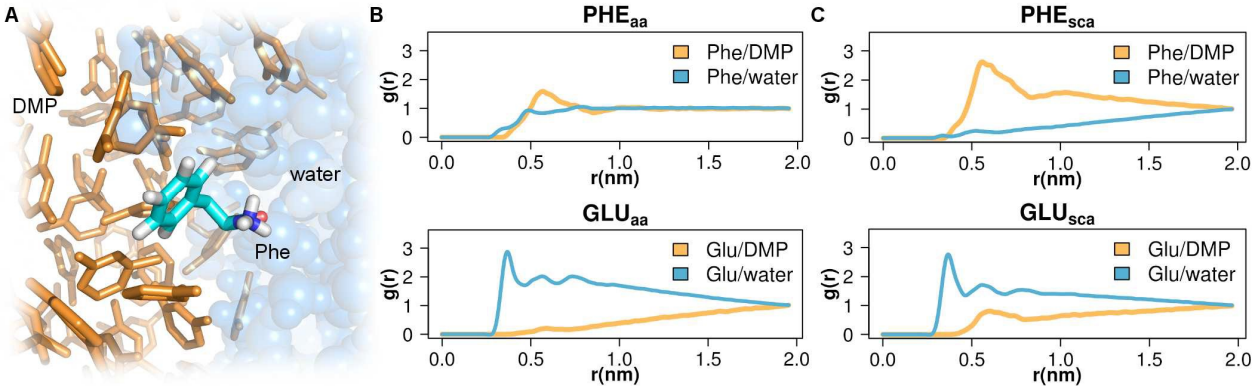


Figure 3

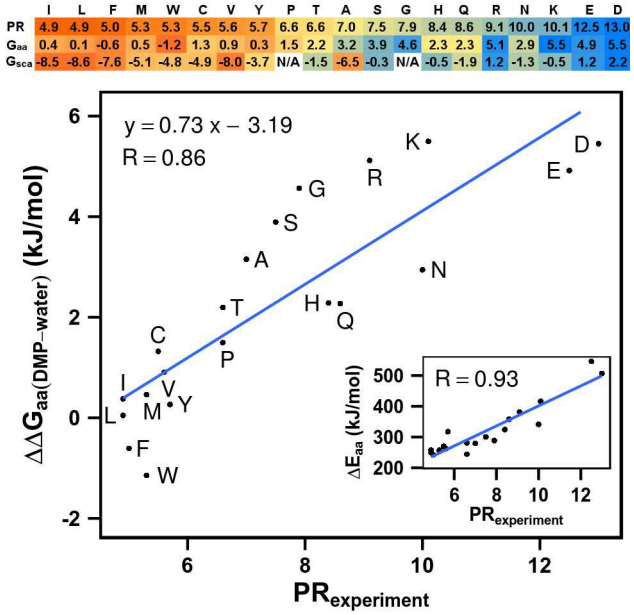


Figure 4

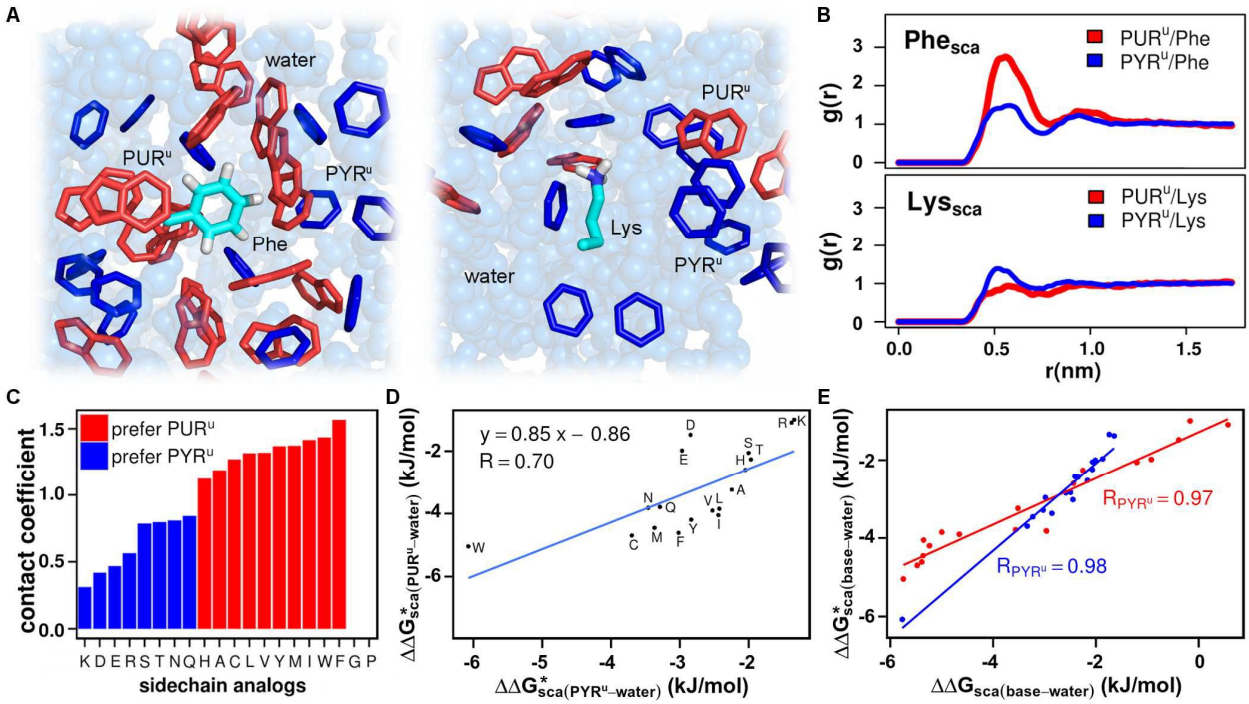


Figure 5