PNA: Protein-Nucleic Acid Complex Structure Prediction

Raymond Zhang

Department of Computer Science and Engineering
Department of Biology
University of Washington

December 16th, 2013

1. Abstract

This paper describes the design and construction of a program called PNA, short for "Protein-Nucleic Acid interactions". PNA takes a protein sequence and one or more nucleic acid sequences, and generates a three-dimensional prediction of the structure of the resulting interaction based on experimentally-derived structures of similar interactions found in the Protein Data Bank. We apply PNA to a well-studied protein-nucleic acid interaction to analyze its efficacy, and suggest possible improvements that can be made to the method.

2. Introduction

Interactions between proteins and nucleic acid (NA) strands are critical to many processes that occur in life, from basic cellular maintenance to the genesis of cancer. Understanding the physical structure of the protein-NA interaction complexes is critical to advancing biological research. Structures of some protein-NA interactions have already been experimentally determined through methods such as x-ray crystallography and NMR imaging; however, some protein-NA interaction structures cannot be determined through these methods due to problems such as failure of the interaction complex to crystallize, and these methods suffer from limitations inherent in all physical experiments and processes—namely, high cost and low speed.

My project, called Protein-Nucleic Acid Complex Prediction, or PNA for short, seeks to overcome the cost and time burdens of traditional experimental methods by using computational techniques to provide an informed prediction of the structure of how proteins and nucleic acid strands interact by exploiting existing knowledge of structures of similar interactions. This paper will describe the principles of PNA as well as describe the details of how PNA operates, from taking protein and NA sequences to producing three-dimensional prediction structures of how they interact.

3. Background

3.a. The Regulog Method

The fundamental idea behind the operation of PNA is the regulog method. The regulog method essentially encapsulates one of the primary observed truths about protein and nucleic acids: similar sequences tend to result in similar structures, and similar structures tend to result in similar interactions. This has shown itself in nature already; for example, closely-related and highly similar protein and NA sequences across different species produce protein-NA interaction families, the interactions of which tend to be very similar in structure to each other. Although simple, this principle is very powerful: it allows us to take our existing knowledge of protein-nucleic acid interactions – most importantly, the experimentally-derived structures of such interactions – and use those to form an educated guess as to how similar protein and nucleic acid sequences would interact. This principle forms the foundation of PNA.

A very similar principle, called the interolog method, applies to protein-protein interactions; the success of the interolog method has already been proven by a previous program by the Samdurala Computational Biology Group, called Protinfo PPC.

3.b. Related Work: Protinfo PPC

Protinfo PPC takes two user-provided protein sequences, and produces a prediction of the structure of the interaction of those sequences.[1] Protinfo PPC is able to predict structures with very high accuracy if templates are found with good sequence similarity; if a template is found with 85% sequence identity to the target proteins, then the structure that Protinfo PPC predicts is less than 5 Å all-atom RMSD (root mean squared deviation) from the actual structure approximately 60% of the time. The success of the interolog method, as used by Protinfo PPC, was a major factor in the decision to base PNA off of the analogous regulog method for protein-NA interactions. The modeling pipeline used by Protinfo PPC, especially the protein-modeling segments of the pipeline, served as a template for PNA.

4. Methods: PNA

PNA is a modeling pipeline composed of distinct steps. As shown in the following figure, the pipeline is divided into the following steps: target specification, template selection, protein alignment generation and structural modeling, NA modeling, model recombination and minimization, and finally model scoring and result selection.

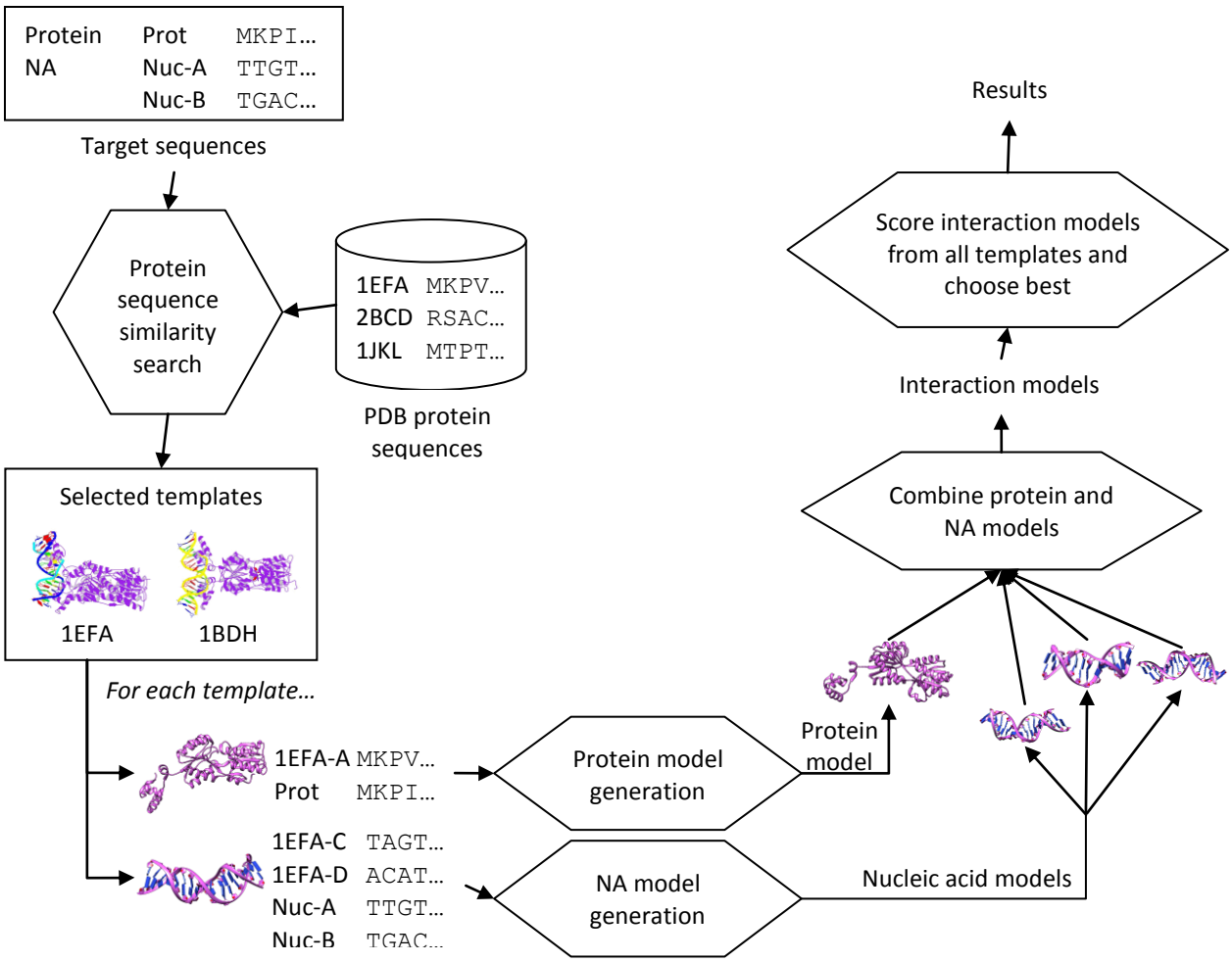

Figure 1: An overview of the PNA pipeline.

PNA represents protein-NA structures via the PDB file format, which describes macromolecule complexes by specifying the 3-D position of each atom in the complex.[2] Atoms in the file are arranged into residues, corresponding to basic assembly units such as amino acids and nucleic acids, which are further arranged into chains; each chain represents a separate molecule. The rest of this thesis will describe manipulations of structures in terms of the effects on the atoms, residues, and chains in the PDB file format.

4.a. Target Specification

The interaction to be modeled, hereafter referred to as the target interaction, is specified by providing one protein sequence and one or more nucleic acid sequences to run PNA on. Each nucleic acid sequence must be marked as being either DNA or RNA; PNA can process a target interaction that contains both DNA and RNA strands. At this time, only one protein sequence per target interaction is supported.

4.b. Template Selection

PNA can either use a provided .pdb file containing a protein-nucleic acid interaction structure as a template, or it can be configured to search through the RCSB Protein Data Bank (PDB), an online repository of structures, for appropriate templates.[3]

4.b.i.Template Provided

If a template structure is given, minor preprocessing is performed on the template: a single interaction structure is chosen if multiple exist within the file, hydrogen atoms are removed from the structure, missing side chains on the protein are rebuilt using standard geometry, and the sequences of the template protein and nucleic acid chains are extracted.

4.b.ii.Template Not Provided

If a template structure is not provided, PNA searches the PDB for protein-NA interaction structures that have protein sequences similar to the target protein sequence. Initial template selection focuses on protein sequence similarity because PDB interaction structures tend to have protein sequences which are much longer than the NA sequences they are involved with; thus, high protein sequence similarity tends to convey a better indicator of overall interaction similarity than comparable nucleic acid structure similarity. Furthermore, major changes to the protein sequence are more likely to result in significant changes to the predicted protein structure than for the NA sequences, and the quality of the predicted structure depends strongly on the similarity of the predicted target structures to those found in the template; this is especially true if the target NA sequences are DNA, because in these cases the DNA strands are often in the form of a double helix, where the replacement of one pair of complementary bases for another pair results in little overall change to the structure.

A set of possible templates is calculated by using the sequence similarity search tools PSI-Blast[4] and SSEARCH[5] over all structures in the PDB that contain both protein and nucleic acids. A protein sequence is considered a possible template if it has a sequence identity of less than 95% and a SSEARCH z-score greater than 120 or a PSI-Blast the z-score over 30. (By default, templates with greater than 95% sequence similarity with the target are rejected; this filter can be disabled, with the caveat that if the user provided a target sequence that exists in the PDB it is likely that a duplicate of that structure will be produced as the result.) This is to ensure that PNA returns interesting results, instead of simply duplicating an existing structure within the PDB.

The top 50 results from the SSEARCH and PSI-Blast searches, as ordered by z-score, are chosen for further consideration. If this results in fewer than 5 templates of low quality, defined as a PSI-BLAST z-score lower than 30, then PNA retries template selection with less stringent z-score requirements in an attempt to find some templates to use; however, the quality of predictions generated using these templates will be lowered.

4.c. Protein Modeling

For the template protein sequences chosen from the previous step, a multiple sequence alignment with the target protein sequence is performed using ClustalW.[6] This alignment specifies how to construct the target protein based on the template protein structure. ClustalW was chosen as the multiple sequence alignment tool for consistency with the protein modeling portion of Protinfo PPC.

For each template interaction structure chosen, the template protein is extracted from the model. The alignment generated between the sequences of the template protein's sequence and the target sequence is then used to perform initial mutation of side chains in the template structure, changing them to the target side chains. After this initial mutation, the rotamer library SCWRL is then used to position the mutated side chains in the model to ensure good packing. At this stage, the templates with the best protein similarity scores are selected for further modeling. (Note that for some alignments this initial side chain mutation and repositioning may fail; only templates for which the previous steps succeeded will be chosen for further processing.)

Each of the remaining template proteins then undergoes loop building to handle insertions and/or deletions as specified by the sequence alignment. Loops are built in order of their occurrence in the sequence for the structure. For short loops (<=5 residues), PNA exhaustively checks all possible configurations based on an n-state phi/psi model and selects the best one as determined by the all-atom protein scoring function RAPDF. [7] For longer loops, PNA applies the segment matching and folding (SEMFOLD) technique to predict the structure of the loop.[8] Finally, the resulting protein structure is minimized using the Energy Calculation and Dynamics (ENCAD) method to produce a prediction of the structure of the target protein sequence.[9] Each template choice results in one target protein structure prediction.

This sequence of steps for modeling the protein portion of the interaction was chosen based on its demonstrated efficacy when used in the modeling pipeline of Protinfo PPC, where it was shown to maintain the critical structures of the binding region so that the predicted interface between proteins remained accurate; this is crucial when the predicted protein model is integrated with the predicted NA model (explained below), where modifying the structure of the binding pocket too much could cause serious steric clashes and ruin the predicted interaction structure.

4.d. NA Modeling

For each template model, PNA models each target NA sequence on all template NA chains using a sliding window approach. For each pair of target NA sequence and template NA chain, if they are of the exact same length then a NA alignment is generated. If the template NA sequence is longer than the target sequence, then every position of the target NA sequence on

the template sequence is attempted, and vice versa if the target sequence is longer than the template.

A sliding window approach was chosen because it would generate target NA models for all possible positions and therefore exhaustively find the best possible alignment from target to template NA. Although this can generate a large number of target NA structure predictions that would have to be combined with the target protein structure predictions, this was deemed an acceptable trade-off largely because the cost of generating the NA models is minimal and the subsequent steps run quickly.

For each alignment of the target sequence on to the template sequence, the target NA structure is constructed nucleotide by nucleotide. Positions on the alignment where the template and target base are identical are directly copied from the template structure.

If the template and target nucleobase are not the same, the nucleobase present on the template structure is replaced with the nucleobase specified by the target sequence; the target nucleobase is taken from a reference nucleotide structure from the PDB. To accomplish this, PNA leverages the fact that nucleobases are constructed from rigid aromatic rings, and so can be effectively represented by rigid rectangular planes attached to the sugar backbone; mutation of a template nucleobase into a target nucleobase can thus be viewed as the translation and rotation of one rectangle, representing the target nucleobase, such that it overlays the current position of another rectangle, representing the template nucleobase. The actual atoms in the structure of the nucleobase remain in the same position relative to each other throughout this movement; to represent this, PNA constructs coordinate frames that represent the position and orientation of the target (source) and template (destination) sites, and uses these to ensure that the nucleobase's atoms are moved to the same relative position in the destination location as they were in the start location. Thus, modeling the nucleobases is broken down into two steps: defining the coordinate frames for the nucleobases such that moving atoms to relative positions in a destination coordinate frame that originally represented a different type of nucleobase makes sense, and finding and applying the appropriate transformation to move the nucleobase's atoms from the source coordinate frame to the destination coordinate frame.

To determine how to define the coordinate frames for the nucleobases, a set of standard nucleotide structures was obtained from the PDB and used as reference models; a coordinate frame was defined for each nucleobase such that the origin corresponds to the nitrogen that connects to the backbone, the –Y axis is along the direction from the attached nitrogen to the backbone C1' atom it bonds to, and the +X direction runs towards the bonding end of the nucleobase. In practice, this means that for purines the origin is placed at N9, the +X axis is defined as running from C8 to C4, and the pseudo-Y axis is in the direction of C8 to N7; for

pyrimidines, the origin is placed at N1, the +X axis is from C6 to C2, and the pseudo-Y axis runs in the direction of N1 to C4. These coordinate frame definitions can be viewed in Figure 2. (PNA ensures the orthogonality of the axes used in the coordinate frame by constructing a Z axis from the X and pseudo-Y axes given and then constructing the true Y axis from the Z and X axes.)

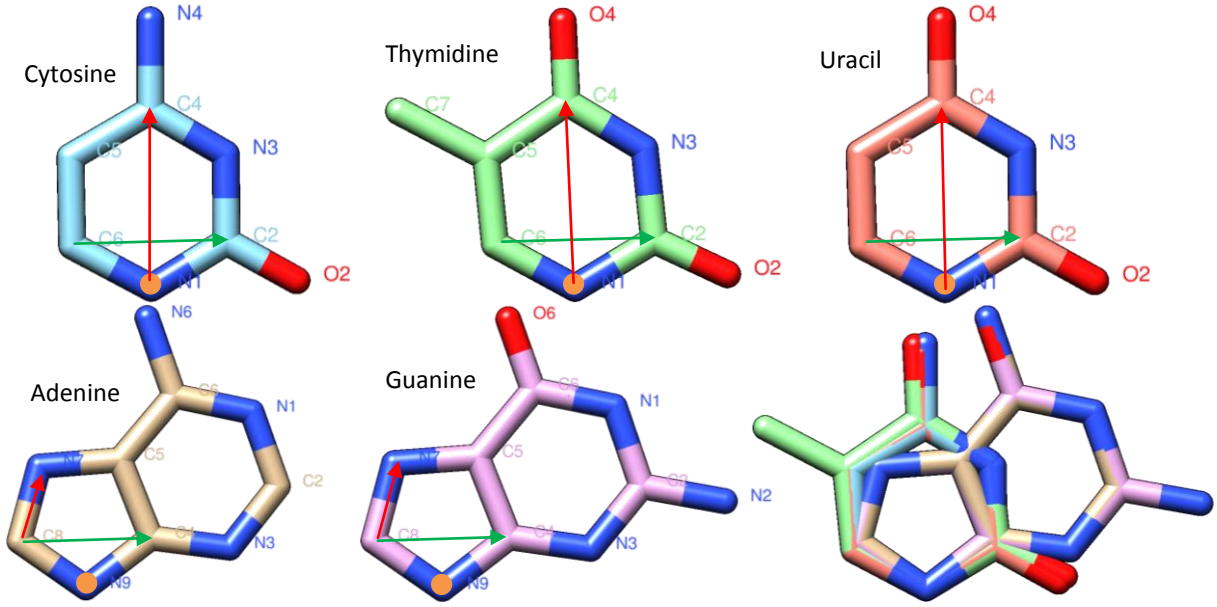

Figure 2: Coordinate frame construction for nucleobases. The origin point is shown in orange, the X axis is shown in green, and the pseudo-Y axis is shown in red. Top row: Cytosine, Thymine, Uracil. Bottom row: Adenine, Guanine, and all five nucleobases superimposed; this last configuration corresponds to how PNA would replace nucleobases in the NA models.

Constructing the coordinate frames in this fashion allows replacement of purines with pyrimidines and vice versa such that they occupy approximately the same space, especially in double helixes; in a double helix, replacing a purine with a pyrimidine on one strand and replacing its opposite-strand counterpart with the complementary base pair results in the nucleobase pair with the same spacing as in the original model, as shown in Figure 3.

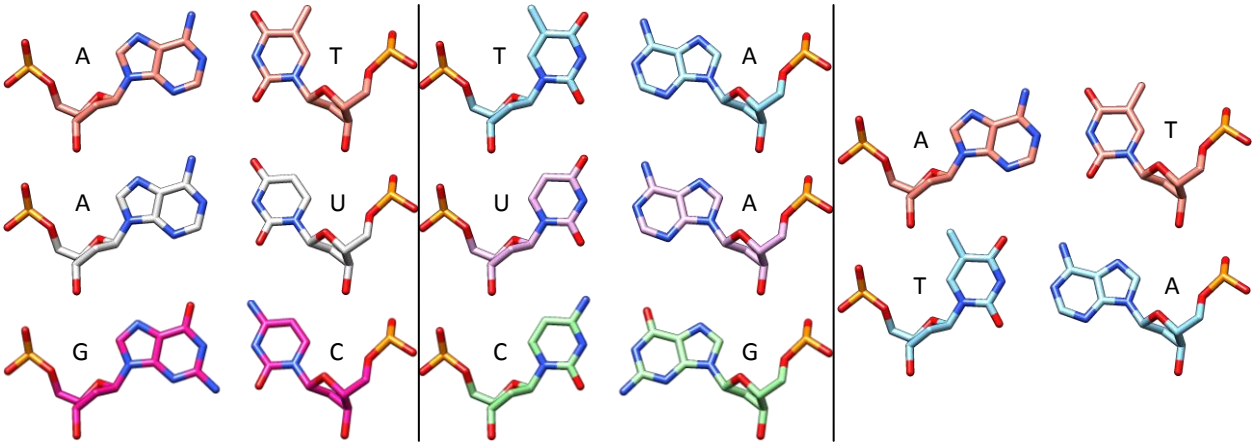

Figure 3: Nucleobase replacement in a base pair from a B-DNA helix. Note how the overall space of the base pairs are equal in all the replacements, even when the purine and pyrimidine sides are swapped. Column 1: AT, AU, CG. Column 2: TA, UA, GC. Column 3: AT, TA.

For each template nucleobase that needs to be replaced, a coordinate frame is constructed for its current position and orientation. The target nucleobase's structure is taken from a reference nucleotide structure from the PDB, and the nucleobase coordinate frame is constructed. Using these coordinate frames, a transformation matrix is calculated that moves the target nucleobase's atoms from their positions in the source coordinate frame to their corresponding places in the destination coordinate frame. This transformation matrix is then applied to all of the target nucleobase's atoms, moving them to their new position, and the template nucleobase's atoms are discarded.

In some cases, a target sequence may be specified as DNA while the template chain is RNA, or vice versa. When the template is RNA but the target is DNA, the O2' atom on the template structure is identified and removed. In cases where the template is DNA but the target is RNA, PNA attempts to add the O2' in the correct position; PNA attempts to achieve this using coordinate frames again. A reference backbone sugar structure from the PDB was obtained and the coordinate frame defined using the same atoms as for DNA and RNA; the reference backbone structure is then moved to the current position of the template, and the O2' from the reference is added to the structure. An example of this is shown in figure 4. Note that only the O2' atom is added to the model; the remaining backbone atoms in the template are used as they are.

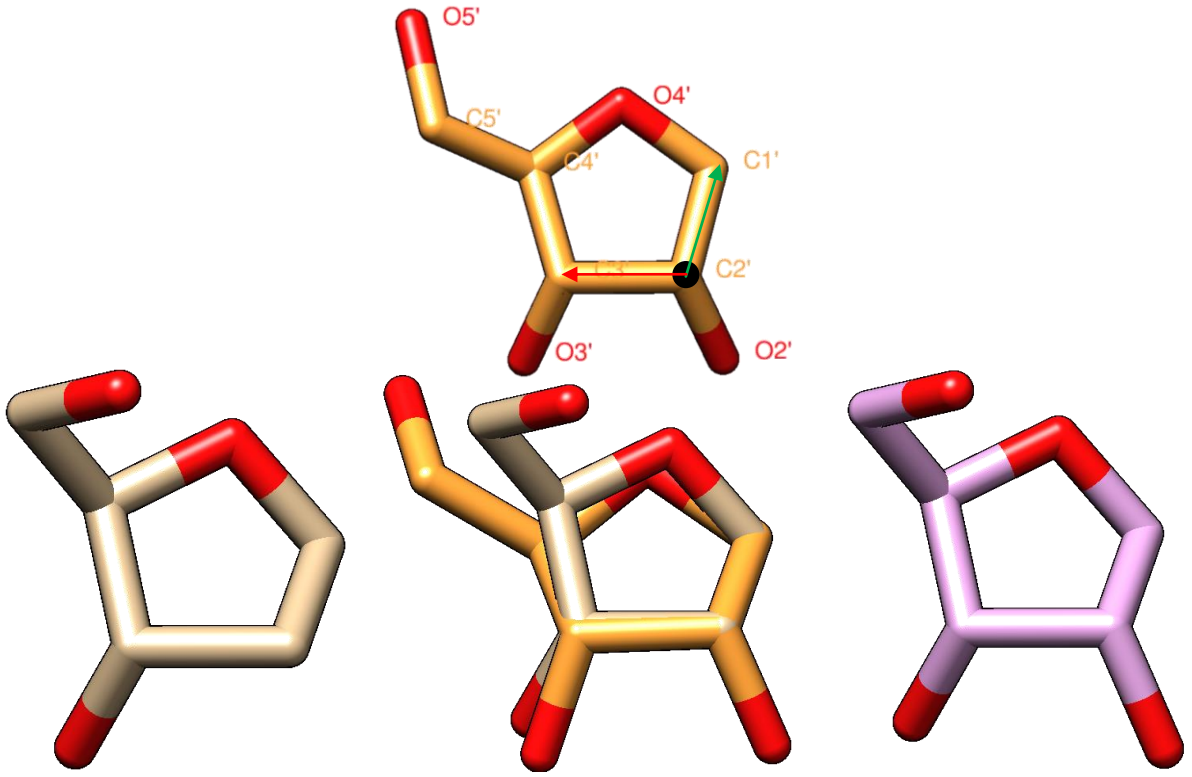

Figure 4: Backbone coordinate frame definition and O2' insertion. Top: An example of coordinate frame definition for a backbone sugar. Origin is in black, X axis is in green, and pseudo-Y axis is in red. Bottom: O2' insertion into a backbone sugar structure, for converting DNA templates into RNA targets. Initial deoxyribose sugar; ribose sugar template superimposed on to the target, and final O2'-containing ribose structure.

4.e. Protein-NA Model Combination, Minimization

The resulting protein and nucleic acid models are overlaid, and the resulting structure undergoes energy minimization using Gromacs with the Amber 99 force field. The Amber 99 force field was chosen specifically for its applicability to both proteins and nucleic acids. This minimization primarily serves to resolve minor steric clashes that result from the changes of amino acids and nucleic acids during target protein and NA structure generation, especially near the binding region(s). In some cases where the steric clashes are too great for energy minimization to compensate, it may result in a "minimized" interaction structure in which the protein and/or nucleic acid structure has become highly deformed, disrupting the interaction; these cases will be detected during the subsequent model scoring, and the affected predictions will be discarded.

4.f. Model Scoring, Result Selection

Scoring is done by using the Samudrala Group's BTTR scoring function to rate the overall protein-NA interaction, RAPDF to evaluate the predicted protein structure, and sequence similarity to determine the quality of the chosen NA alignment windows.[10] The scores from the three functions are combined using a weighted rank average that favors the ranking produced

by BTTR; the best scoring models across all templates are then given back as the final results. BTTR was specifically chosen to evaluate the overall interaction due to its strong displayed discriminatory power for protein-DNA complexes, which is integral in allowing PNA to select the best predictions to serve to the user.

5. Case Study: Lac Repressor / Operon Interaction

To determine the effectiveness of PNA, we applied PNA to a well-studied protein-NA interaction: that between the *E. coli* lac repressor and the lac operon. The lac repressor/operon system was chosen because of the wealth of experimental data on the interaction properties of various repressor/operon mutant combinations.

5.a. Methods

We chose to run PNA on a set of repressor / operon mutant combinations for which binding data was known. In a previous paper, Milk, Daber, and Lewis studied the interaction of various repressor/operon pairs; they reported a group of repressor-operon pairs with a spectrum of repression ratios, calculated as the amount of expression of operon in the presence of repressor divided by the expression of lone operon, which serves as indicators of the quality of the repressor-operon interaction.[11] Based on their data, we chose to run PNA on a set of test cases consisting of 5 repressor-operon pairs with low repression ratios (indicating good binding) and 5 repressor-operon pairs with higher repression ratios (indicating poor binding), as indicated in Table 1, with the intent of showing how well PNA can discriminate between the good and bad interaction pairs.

ASR-GGT	0.008	ATR-GGA	0.018	HTN-TTT	0.021	STN-TTA	0.289	TGN-TTA	0.325
AAR-GGA	0.016	TSR-GGT	0.019	AKN-TAC	0.241	TSA-CGT	0.301	VAN-TTA	0.383

Table 1: Test cases used for PNA, with corresponding repression ratios. The first three characters of each name indicate the residues at positions 17, 18, and 22 of the lac repressor.[12] The last three characters of each name indicate the nucleotides at positions L6, L5, and L4 of the symmetric lac operon.[13] The two NA sequences used for each test case were L7 – L1 and R1 – R7 from the symmetric lac operon. These ten cases were the five repressor/operon pairs with highest repression ratios and the five with lowest repression ratios given in Table 1 of the Milk, Daber, and Lewis paper.

For each of these test cases, PNA was provided with the repressor mutant sequence and the sequences of the two operon strands. PNA was configured to find the five best template models, generate predictions based on those models, and return the top five models per template. For this particular experiment, PNA was allowed to use templates with greater than 95% protein sequence similarity to the target protein sequence, because the test cases varied only in the selection of protein and NA residues in the interaction region.

5.b. Data

Because the target protein sequences were very similar across all test cases, PNA chose the same five protein templates for all test cases. These protein templates, in form PDB code – template chain ID, were 1EFA-A, 1EFA-B, 1JWL-B, 2PE5-A, and 2PE5-B. For each of these template selections, we gathered the resolution and R-value from the PDB entry, and obtained the BTTR scoring of the protein-NA interaction in the template structure; this data is shown in table 2.

Template code – protein chain ID	Structure resolution (Å)	R-value	BTTR score
1EFA-A	2.6	0.247	-26.17807
1EFA-B			-25.00444
1JWL-B	4	0.249	29.81626
2PE5-A	3.5	0.236	-65.72916
2PE5-B			-64.1931

Table 2: Structure resolution, R-value, and BTTR scores for each template selection. Structure resolution and R-values were obtained from the PDB entries for each template code. The BTTR score was calculated for the interaction between the protein chain and the two whole nucleic acid strands that the chain interacted with; a lower BTTR score indicates a better quality template. The BTTR scores indicate that 2PE5-A and 2PE5-B are the highest-quality templates, with 1JWL-B as the worst template and 1EFA-A and 1EFA-B in the middle.

For each template – test case combination, PNA generated a group of predictions; we selected the top 1, 3, and 5 (as reported by PNA) interaction models for each combination, and scored them using BTTR. We used the BTTR scores to rank the results from each template across all test cases, and calculated both the raw correlation and rank correlation between the BTTR scores and the repression values. In addition, we used the BTTR scores to assign each result from a template into one of two groups based on whether its BTTR score was better or worse than the median. Because the test were are divided into five low-repression-ratio cases and five high-repression-ratio values, we counted the number of results for low-repression-ratio cases that appeared in the group with better BTTR scores and the number of results for high-repression-ratio cases that appeared in the group with worse BTTR scores, and determined the probability that a sorting that had this many or more correct selections would have occurred by chance. The data obtained is shown in Table 3.

	1EFA-A	1EFA-B	1JWL-B	2PE5-A	2PE5-B
Top 1 (10 results)	-40.07 ± 5.8401 -0.1236 (0.6331) -0.3697 (0.8534) 4 (0.8281)	-41.41 ± 5.3442 0.1729 (0.3164) -0.1152 (0.3758) 4 (0.8281)	-2.68 ± 3.4776 -0.3013 (0.8011) -0.4061 (0.8778) 4 (0.8281)	-47.05 ± 3.3312 0.8778 (0.0004) 0.6970 (0.0126) 8 (0.0547)	-48.90 ± 5.3019 0.0997 (0.3920) -0.1879 (0.6983) 6 (0.3770)
Top 3 (30 results)	-32.33 ± 7.2151 0.1127 (0.2765) 0.1875 (0.1605) 20 (0.0494)	-36.61 ± 5.7259 0.2087 (0.1342) 0.1128 (0.2764) 16 (0.4278)	-0.85 ± 2.7879 -0.1352 (0.7619) -0.1720 (0.8181) 16 (0.4278)	-44.81 ± 3.5389 0.3523 (0.0281) 0.2004 (0.1441) 20 (0.0494)	-47.02 ± 4.7275 0.3047 (0.0508) 0.1786 (0.1725) 18 (0.1808)
Top 5 (50 results)	-29.623 ± 6.8215 0.1285 (0.1869) 0.1849 (0.0993) 32 (0.0325)	-33.74 ± 5.8975 0.0893 (0.2687) -0.0029 (0.5080) 32 (0.0325)	0.66 ± 3.5177 0.0540 (0.3549) -0.0518 (0.6396) 26 (0.4439)	-43.61 ± 3.4593 0.2792 (0.0248) 0.1022 (0.2400) 30 (0.1013)	-45.30 ± 4.7084 0.3143 (0.0131) 0.2408 (0.0460) 34 (0.0077)

Table 3: Results of applying PNA to the lac repressor / operon test cases. The top row lists the protein templates used by PNA. The top 1, 3, and 5 results for each template were combined across all test cases. For each cell, the first line is the mean and standard deviation of BTTR scores, the second line is the raw correlation between repression values and BTTR scores with associated p-value, the third line is the rank correlation of repression values and BTTR scores with associated p-value, and the last line is the number of correctly-sorted results and associated p-value. For correlations, the p-value is the one-tailed probability of obtaining a higher correlation by chance. The p-values for the number of correctly sorted results are the probability of obtaining that many or more correctly-sorted values by chance. This data indicates that PNA generates results that parallel expected data, but there is dependence on the quality of the templates selected and that the final scoring step used by PNA to rank the generated results could be improved.

5.c. Analysis

The templates chosen by PNA are structures of the lac repressor bound to the lac operon; this is as expected, because we allowed selection of template structures with protein sequences very close to the target sequence. By examining the BTTR scorings of the template choices, we can see that 2PE5-A and 2PE5-B are the best templates, 1JWL-B is the worst, and 1EFA-A and 1EFA-B are in the middle. Note that this corresponds to the relative R-values of the templates.

In comparing the number of correctly-sorted results and associated p-values obtained from the different templates, it is apparent that the quality of the results generated by PNA is highly dependent on the template selected; for example, 1JWL-B resulted in poor sortings for the top 1, 3, and 5 results. The poor sorting produced using 1JWL-B as the template is likely due to the overall poor quality of those results, as indicated by the average BTTR score for those results, which can be traced back to the poor quality of the template. This indicates that initial template selection could be further refined, perhaps by using BTTR as another means to rank and select templates.

Comparing the number of correctly-sorted values obtained from taking the top 1, 3, and 5 results per template, it generally appears that PNA performs better if taking into account more

of the final models generated instead of less, especially with regards to the jump from including only the top 1 result to including the top 3 results; this suggests that the final scoring and ranking of models generated by PNA is not as fine-tuned as might be desired, and should be updated to better discriminate among the results. However, for each template selection other than 1JWL-B, the top 3 and/or top 5 results generate a sorting with p-value less than 0.05, which indicates that PNA is generating models that are in line with the quality of the inputs given; in addition, note that the sorting generated for the top 1 result for each test case from template 2PE5-A had 8 correct selections out of 10, with p-value 0.054688, which is an example of a case where PNA's final scoring and ranking was able to pick the correct top result for each test case.

Overall, it seems that PNA can generate results that parallel experimental results, as indicated by the good sorting of the top 3 and/or 5 results for templates 1EFA-A, 1EFA-B, 2PE5-A, and 2PE5-B; however, the accuracy of PNA is dependent on the template selection, as seen with 1JWL-B, and the final scoring and ranking of results generated by PNA may need to be updated to properly discriminate between the top results.

6. Discussion

6.a. Limitations

One case in which PNA will produce less-than-optimal results is if the only templates that can be found come from a multimeric structure where multiple protein subunits come in contact with a single NA chain. This will result in only a single protein chain being modeled along with the NA strand, and so part of the NA strand that used to be in contact with other subunits will now be dangling free. This is even more pronounced if the interaction with the missing protein subunit would have significantly deformed the structure of the NA strand. If the missing protein subunits are critical for the overall interaction between the protein and NA strands, the resulting predictions will be worse than usual. Some correction is made for this in the use of BTTR in the final scoring and model selection; because BTTR evaluates the structure as a whole, it is able to give a worse score to those predictions in which the interaction between protein and NA strands has been disrupted due to the actions of PNA.

Currently, template interaction models are primarily chosen based on the sequence similarity of the protein involved with the target protein sequence. This was considered acceptable due to the fact that normally protein structures are far less tolerant of amino acid changes both in the binding pocket and in the rest of the structure than nucleic acid structures, and that generally the nucleic acid strands found in the interaction structures in the PDB tend to be far shorter than the protein involved. This leaves room for improvement, however; future work could be directed in using the similarity of the target NA sequence to those involved in the

interaction structure to better select templates, especially with regards to the NA bases that are directly involved in the interaction with the protein.

Currently, the protein and nucleic acid structures are modeled independently; the protein is modeled without knowledge of the nucleic acids involved, and likewise the nucleic acid strands are modeled without knowledge of the protein. This is sufficient to generate reasonably good models, as the nucleic acid modeling does not change the position of the backbone, and the protein modeling (due to the sequence similarity) usually does not change the template structure very much; once the models are recombined, the short energy minimization stage serves to rectify most minor issues that arise. However, this can sometimes lead to cases where the interaction between the protein's binding pocket and the nucleic acid strand is ruined due to the mutation of the template protein and NA structures causing steric clashes. A possible solution to this would be to merge protein and nucleic acid alignment generation such that sequence alignments that lead to clashes in the binding region are discarded, as was done for Protinfo PPC.

6.b. Future Work

We hope to make PNA available as a web server online for public use in the near future, under the umbrella of the Protinfo collection of web servers created by the Samudrala Computational Biology Group.

7. Conclusion

This paper has described the construction of PNA, a program based on the regulog method that predicts the structures of interactions of protein and nucleic acid sequences. PNA takes protein and nucleic acid sequences, and uses known structures of similar interactions as templates to predict the structure of the target interaction; these templates can either be provided by the user, or found from a search of the Protein Data Bank. The process of constructing the target structure based on the template involves modeling the involved protein and nucleic acid strands separately on their templates and subsequently recombining the models to generate a prediction; the best predictions across all templates are presented as results to the user. An application of PNA to a set of lac repressor mutant and lac operon mutant interactions demonstrated that PNA was able to generate results that paralleled the experimental data, but that the final scoring and ranking step of PNA could be further improved, and that the results of PNA are highly dependent on the quality of the template structure(s) selected. We hope to have PNA running as a web server in the near future for public use.

8. Acknowledgements

I thank Mark Minie, who helped me find an appropriate case study for this thesis, Thomas Wood, whose usage of my NA-modeling program drove me to refine the method which I was using, and Gaurav Chopra, who particularly helped me with using energy minimization tools for PNA. I would like to thank Jeremy Horst, whose assistance was exceedingly helpful when I was just starting in the research group. I would like to thank all the members of the Samudrala Computational Biology Group for their invaluable support while I was working on this project. I would like to thank Martin Tompa, my CSE research adviser for this project, for his help with refining this thesis. And, of course, I would like to express my sincerest gratitude to Ram Samudrala, for his mentorship and guidance during my years in his group; he is the main force that helped me grow from a young, raw college student into the researcher that I am today.

Molecular graphics were generated using the UCSF Chimera package. [14] Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). Protein-NA complex images for Figure 1 were taken from PDB entries for 1EFA and 1BDH. [15,16]

This work was supported in part by a Barry M. Goldwater Scholarship, a Mary Gates Research Scholarship, a Microsoft Endowed Scholarship, and a Kildall Endowed Scholarship to Raymond Zhang. This work is supported in part by a 2010 US NIH Director's Pioneer Award (DP1OD006779) and NSF CAREER Award (0448502) to Ram Samudrala.

9. References

[1] Kittichotirat W, Guerquin M, Bumgarner RE, Samudrala R. Protinfo PPC: A web server for atomic level prediction of protein complexes. Nucleic Acids Research 2009;37: W519-25.
[2] PDB file format documentation is available at http://www.wwpdb.org/docs.html
[3] RSCB PDB is located at http://www.rcsb.org. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne The Protein Data Bank. Nucleic Acids Research 2000;28: 235-242.
[4] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402.
[5] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc. Natl Acad. Sci. USA. 1988;85:2444–2448.
[6] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–2948.
[7] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 1998;275:895–916.
[8] Samudrala R, Levitt M. A comprehensive analysis of 40 blind protein structure predictions. BMC Struct. Biol. 2002;2:3.
[9] Levitt M, Hirshberg M, Sharon R, Daggett V. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phy. Commun. 1995;91:215.
[10] Bernard B, Samudrala R. A generalized knowledge-based discriminatory function for biomolecular interactions. Proteins: Structure, Function, and Bioinformatics 2009; 76: 115-128.

[11] Milk L., Daber R., Lewis M. Functional rules for lac repressor-operator associations and implications for protein-DNA interactions. Protein Sci. 2010; 19: 1162–1172.
[12] For this test, the base lac repressor sequence was that of the *E. coli* lac repressor, taken from http://www.uniprot.org/uniprot/P03023
[13] Simons A, Tils D, von Wilcken-Bergmann B, Muller-Hill B. Possible ideal lac operator: Escherichia coli lac operator-like sequences from eukaryotic genomes lack the central G X C pair. Proc. Natl. Acad. Sci. USA. 1984;81(6):1624–1628.
[14] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.
[15] Image from the RSCB PDB of PDB ID 1EFA. Bell CE, Lewis M. A closer view of the conformation of the Lac repressor bound to operator. Nat. Struct. Biol. 2000;7:209-214.
[16] Image from the RSCB PDB of PDB ID 1BDH. Glasfeld A, Schumacher MA, Choi KY, Zalkin H, Brennan RG. A Positively Charged Residue Bound in the Minor Groove Does not Alter the Bending of a DNA Duplex. J. Am. Chem. Soc. 1996;118:13073-13074.