The Pennsylvania State University

The Graduate School

The Eberly College of Science

STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF THE TRANSCRIPTION FACTOR PDX1 AND ITS INTERACTIONS WITH DNA AND PROTEINS

A Dissertation in

Chemistry

by

Monique Bastidas

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

August 2015

The dissertation of Monique Bastidas was reviewed and approved* by the following:

Scott A. Showalter
Associate Professor of Chemistry
Dissertation Advisor
Chair of Committee

Philip C. Bevilacqua
Professor of Chemistry

Christine Keating
Professor of Chemistry

David Gilmour
Professor of Molecular and Cell Biology

Barbara Garrison
Shapiro Professor of Chemistry
Head of the Department of Chemistry

*Signatures are on file in the Graduate School

ABSTRACT

Homeobox proteins are vital DNA-binding transcription factors that control the spatial and temporal expression of cells and tissue during early development. These proteins are necessary for other biological processes, including the activation of genes. The DNA-binding homeodomain motif of homeobox proteins bind sequence specifically to double stranded DNA containing the core binding site (5′-TAAT-3′), achieving high specificity *in vivo*. Our current understanding of the mechanism by which these transcription factors recognize their cognate sites is not clear. Moreover, the homeodomain motif is often a small domain within a larger polypeptide chain tethered to intrinsically disordered domains at one or both termini. Intrinsically disordered proteins (IDPs) play a vital role in maintaining cellular functions. The homeobox transcription factor, Pdx1 is an IDP, important for maintaining proper function of the pancreas and β-cells by recognizing A-box elements (TA-rich sequences containing a 5′-TAAT-3′ core site). The overarching goal of this dissertation was to further the understanding of the mechanism by which homeobox proteins recognize their target sites and bind with high specificity, as well as illuminate the structural and functional role of the disordered C-terminus of Pdx1.

Using isothermal titration calorimetry (ITC), it was shown that the homeodomain (HD) of Pdx1 exhibited varying binding affinities for promoter-derived (islet amyloid polypeptide (IAPP) and insulin) and consensus-derived DNA sequences indicating nucleotide sequence discrimination in the flanking regions. The calorimetric data demonstrated that Pdx1 preferentially bound sequence specific insulin elements containing a pentameric DNA sequence (5′-CTAAT-3′) rather than IAPP elements encompassing a 5′-TTAAT-3′ sequence. Furthermore, binding studies of the HD with a panel of consensus-derived mutants demonstrate a slight preference for an insulin-like 5′-CTAAT-3′ sequence over an IAPP-like 5′-TTAAT-3′ sequence based on a combination of a tighter binding constant and a more favorable binding enthalpy.

Molecular dynamics simulations further showed a strictly conserved arginine has preference for the pentanucleotide sequence 5′-CTAAT-3′ rather than 5′-TTAAT-3′, due to a more electronegative binding pocket. To determine the residue(s) involved in sequence specificity upstream from the core site and understand the role of residues in the C-terminus in DNA-binding, a slightly longer homeomdomain construct (HDx) was studied using nuclear magnetic resonance (NMR) and ITC. The construct was much more stable than the HD construct, which was likely a result of the increased stability of helix 3 as evidenced in the NMR data. Using ITC, the binding affinities of HDx to a panel of DNA were in the low nanomolar range and tighter than the HD affinities to the same panel of sequences. In an attempt to identify the residues responsible for imparting specificity upstream from the core site, a mutational analysis with HDx was carried out. Neither mutations resulted in decreased specificity; however, one mutation contributed slightly to the binding enthalpy.

Many folded proteins contain regions that are random coil in structure or form transient secondary structures, which are important for cell signaling, molecular recognition, and transcription. The non-homeodomain regions of Pdx1, which make up a large portion of the protein, are predicted to be disordered based on the amino acid sequence. However, there is no concrete evidence to suggest whether the non-homeodomain regions of Pdx1 are truly unstructured or whether they adopt residual structure. The NMR data of the Pdx1 C-terminus shows that this region adopts random coil structure in solution. Biologically, the plasticity of the intrinsically disordered regions is utilized as a probe to weakly associate with multiple binding partners. One such partner for Pdx1 is SPOP, which is involved in mediating the interaction between an ubiquitin ligase complex and its substrate. Calorimetric data and NMR demonstrated a direct binding interaction and identification of the SPOP binding regions of Pdx1 by NMR.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would not have completed this educational journey without the support of numerous people and programs. I will be forever grateful to all those who have helped me along the way, you all have shaped my journey in one way or another, which has landed me where I am today. Although I have only listed a select few to acknowledge, you know who you are and I am infinitely grateful. This dissertation is specifically dedicated to my mother, brother, maternal grandparents, husband, women and underrepresented communities who may use this as a source of inspiration for academic endurance.

First and foremost, I would like to thank my families, the Gonzales' and Bastidas', for instilling in me a constant thirst for success. Without their support and high expectations of me, I would not have pursued graduate training nor completed this dissertation. I would like to personally thank my mother, brother, and maternal grandparents for the continual care and support, and my mother for giving me pep talks and always sending me care packages.

Beginning with the undergraduate support I received, I would like to thank the Science Educational Equity (SEE) program, especially Pamela King and Dr. Mary McCarthy-Hintz, at California State University, Sacramento for introducing me to the many underrepresented research programs that I participated in, which led me to understand my passion for research and want to pursue higher education. Because of the SEE program, I was led to the great mentorship of Drs. Thomas R. Peavy and Roy W. Dixon, who helped shape me into the researcher I am today. In all sincerity, I am truly grateful for the constant constructive criticism that enabled me to remain positive during criticisms and failures endured during my undergraduate and graduate career.

I would like to thank my committee members, Drs. Philip C. Bevilacqua, Christine Keating, and David Gilmour, for their guidance during the first few years in the graduate Chemistry program. A special thanks goes to Dr. Scott A. Showalter for accepting me into his laboratory to work on a wonderful project that is difficult for me to let go and for pushing me to meet, head on, the goals I set out for myself. My graduate career at Penn State could not have been a success without the friendship and support of Brittany Nagle, Chelsea Hull and from members of the Showalter lab, in particular Dr. Christopher Wostenberg, Dr. Chad Lawrence, Dr. Kaycee Quarles, and Roderico Acevedo.

Lastly, I would like to acknowledge and considerably thank my husband, Dr. Adriano Z. Zambom, who has remained the most supportive in seeing through the completion of my dissertation during that last five years. Thank you for translating your patience and calm attitude, and keeping me sane by convincing me to enjoy life every once in a while during graduate school. Despite the trials and tribulations we endured as a result of intra-continental distance, our relationship has become stronger. Now, together we will travel the world and enjoy what life has to offer.

Chapter 1

Introduction

Abstract

Broadly, the structural and functional roles of homeodomain proteins, emphasizing their description as intrinsically disordered proteins, are described herein. For example, the pancreatic duodenal homeobox 1 (Pdx1) is necessary for multiple biological functions within the early developmental stages and regulation of peptide hormones expressed in the pancreas in adults. Although much of the DNA-binding role of Pdx1 is well established, the mechanism by which Pdx1 regulates pancreatic-specific genes and the structure and functional role of the regions outside the homeodomain are less clear. This dissertation aims to provide insight into the mechanism of Pdx1-mediated gene expression by employing spectroscopic and calorimetric methods, in particular nuclear magnetic resonance (NMR) and isothermal titration calorimetry (ITC). Finally, a summary of the dissertation chapters is provided to close this introduction.

1.1 Homeodomain

In higher organisms, anatomical development occurs in a spatial and temporal manner, which is controlled by the master regulatory homeotic genes.[1-3] These genes encode a set of homeobox proteins whose functions are to regulate transcription by sequence-specific DNA interactions. Encoded within a homeobox gene is a 60 amino acid homeodomain motif responsible for the sequence-specific recognition of target genes. By regulating gene expression,

homeodomain proteins function in multiple roles, such as specifying the morphological development of multicellular organisms, cell differentiation, and cell maintenance.[1, 3]

The homeodomain is composed of 60 residues, which can be referenced in the canonical numbering (i.e., 1-60) or amino acid sequence numbering. Throughout this chapter the canonical numbering of the homeodomain will be used (Figure 1-1A). In addition, descriptions of the dsDNA, core sequence, and flanking sequence will be heavily referenced in this chapter. Therefore, in order to orient the reader the directionality of the promoter elements relative to the transcription start site in a hypothetical promoter are defined in Figure 1-1C.

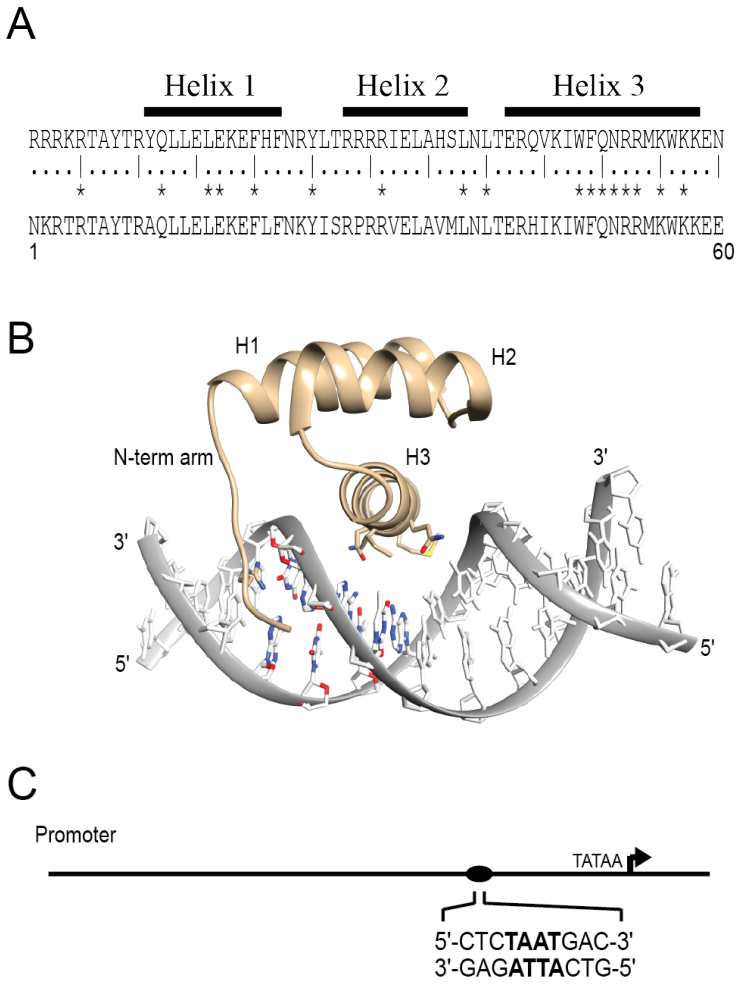

Figure **1-1.** Canonical homeodomain sequence and structure. (A) The consensus amino acid sequence (aa 1-60) derived from 346 homeodomains (upper) and the Pdx1 homeodomain sequence (lower).[1] Highly conserved amino acids are marked by asterisks. (B) Crystal structure of the DNA-bound configuration of the homeodomain motif displaying the disordered N-terminal arm and the three helices (pdb ID: 2H1K). (C) Representation of a promoter region, where the directionality of the binding element is defined relative to the transcription start site.

Homeobox proteins share a consensus sequence (Figure 1-1A), in which a series of conserved amino acids in helix 3 and the N-terminal arm enables the sequence-dependent recognition of dsDNA, shown in Figure 1-1B. Variations in the conserved amino acids alter the sequence selectivity of the homeodomain. In 1993, the DNA-bound solution structure of the

Antennapedia (Antp) homeodomain of *Drosophila melanogaster* was established by nuclear magnetic resonance (NMR).[4, 5] The solution NMR structure of the homeodomain revealed an all α-helical motif, where the first two helices are anti-parallel to each other and helix 3 is aligned perpendicular to the plane of helix 1 and 2. The recognition helix, helix 3, is positioned well within the major groove of double stranded DNA (dsDNA) where sequence-specific interactions are made. Namely, highly conserved residues Ile47, Gln50, Asn51, and Met54, in many homeobox proteins, participate in DNA recognition. An additional and critical component of the homeodomain motif is a relatively short disordered arm at the N-terminus. In the N-terminal arm of the homeodomain, two positively charged residues Arg/Lys3 and the strictly conserved Arg5 contribute to binding affinity. Based on crystallographic evidence, Arg5 inserts into the minor groove to interact with the first 2 bases in the tetranucleotide core sequence, in this case Thy and Ade in the core.[2, 6] An illustration of the structure of a homeodomain and its interactions with DNA are shown in Figure 1-1B.

The DNA-binding function of many homeodomain proteins has been well established in the literature, leading to a generally conserved mechanism. From the deposited structures of homeodomain-DNA complexes, the sequence specific contacts made by residues with the tetranucleotide core sequence, and the 3′ proximal DNA sequence are easily identified. However, the sequence specificity of homeodomains towards the 5′-region from the core binding site has not been as well characterized as the 3′-region. It is well known that direct sequence readout of bases, due to distinct hydrogen bonding patterns, occurs in the major groove, but can also take place in the minor groove.[7-11] An alternative mechanism to DNA recognition is indirect readout of DNA, in particular the minor groove, whereby electrostatics and groove shape are used to distinguish DNA sequence.[12] For example, the strictly conserved arginine in position 5 in the N-terminal arm of homeodomain proteins is responsible for minor groove interactions. While there is some sequence conservation in the N-terminal arm of homeodomains, there is also significant

variation. What is the role of the other charged residues in the N-terminal arm in sequence recognition and DNA-binding?

One homeotic protein, Pdx1 in particular, is implicated in a diabetic phenotype and is introduced in Chapter 1.3. Pdx1 is an excellent model system for studying homeodomain proteins because it represents a specific class of homeobox proteins and intrinsically disordered proteins. Interestingly, homeodomain proteins are often thought of as small well-folded proteins, however, their polypeptide chains are greater than 60 residues in length. Bioinformatics and an increasing number of structural studies have revealed the non-homeodomain regions of the polypeptide chain are structurally dissimilar to the homeodomain in the sense that they are intrinsically (i.e., natively) disordered. For transcription factors and homeodomain proteins, a large degree of their primary sequence adopts a disordered state, making them ideal candidates to investigate the potentially advantageous biological role of native protein disorder.

1.2 Intrinsically Disordered Proteins

Homeodomains are often embedded within a larger protein containing regions of disorder at the N-terminus, C-terminus, or both. Despite the current view of the central dogma of structural biology, intrinsically disordered regions (also known as intrinsically disordered proteins, or IDPs), which are often found in the same polypeptide chains as well-folded domains, participate directly in important cellular functions. For example, IDPs are involved in cell signaling, molecular recognition, and regulation of transcription.[13-15] Due to their relatively recent discovery, and their divergent primary sequence and structural characteristics with respect to those found in cooperatively folded domains, the structure-function relationships of IDPs are only beginning to be uncovered. This leads to the importance of fully characterizing disordered regions

in order to understand their mechanistic role in protein function, which has been a central objective of this dissertation.

Structurally, IDPs lack unique and/or temporally stable tertiary structure and often show little to no secondary structure. However, secondary structures may be induced upon binding to partners, although this is not universally necessary in order to define specificity. Some well documented cases of binding-induced order include the interaction of pKID/KIX, which is representative of local folding of disordered transcription activation domains upon interaction with other proteins and induced folding upon DNA-binding exemplified by the interaction of the high mobility group transcription factor LEF-1 with target DNA.[16,17]

Over the past 15 years, various structure prediction algorithms have been developed to predict the degree of disorder based on the amino acid sequence, which has worked well because non-ordering sequences are quite distinct from those of proteins that fold cooperatively. Notably, IDPs are enriched in charged and/or polar residues, such as the amino acids M, K, R, S, Q, P, and E, while lacking non-polar amino acids that tend to form hydrophobic cores persistent in well-folded proteins.[13-15, 18] The data collected from these disorder predictors shows between 35-51% of the amino acid sequences in Eukaryotic proteins are intrinsically disordered (i.e., the proteome is up to 50% disordered).[14] Although predictors are an excellent starting tool to determining intrinsic disordered in a particular system, a quantitative method for structure determination of IDPs is vital to establish a clear structure-function relationship.

As evidenced in the examples described above, transcription factors exhibit a great degree of disorder; the predictors mentioned above indicate that nearly 70% of the polypeptide chains from transcription factor proteins can be categorized as regions of disorder.[13] Moreover, disorder is not only found in eukaryotes; prokaryotes have also been predicted to have a reasonable degree of disordered regions, although the degree of disorder in proteins increases with increasing hierarchy in organisms.[13] The increased prevalence of intrinsic disorder in

eukaryotes is theorized to be evolutionarily favorable in increasing the functional complexity of proteins.[19, 20] Several types of transcription factors display greater degrees of disorder than others. To list a few, these include the DNA-binding type: A-T hook, HMG box, and homeobox, and the Zinc-finger: C2H2-type.[13] One DNA-binding transcription factor containing the homeobox motif is described further and is selected as the model protein for transcription factors, homeodomain proteins, and IDPs.

1.3 Pancreatic Duodenal Homeobox 1

Pancreatic duodenal homeobox 1 (Pdx1) is a well-conserved homeobox protein involved in the development of the pancreas and duodenum, cell differentiation of pancreatic β-cells, and ensuring proper function of the mature endocrine pancreas.[21, 22] Following cell differentiation and maturation of the pancreas, Pdx1 expression is primarily restricted to β-cells, while low levels of Pdx1 are expressed in other Islet of Langerhans (α- and δ-cell types) and neural cells. Pdx1 maintains glucose homeostasis by regulating transcription of insulin and other β-cell specific genes, such as *glucokinase*[23, 24], *islet-amyloid polypeptide*[25] (*IAPP)* and *glucose transporter type 2*[26] *(GLUT2)* genes. By extension of the TAAT sequence preference common to all homeobox proteins, Pdx1 binds in a sequence specific manner to dsDNA regions containing the core binding site 5′-TAAT-3′. Table 1-1 displays the Pdx1 promoter elements for the β-cell specific genes mentioned above, whereas Figure 1-2 illustrates the promoter structures for these genes.

Table **1-1.** Sequence alignment of Pdx1 A-box elements from human promoters. Cognate binding elements from the *insulin, islet amyloid polypeptide* (*IAPP*), *glucokinase*, and *glucose transporter 2 (GLUT2)* promoters displaying the core Pdx1 binding site.

Gene	Sequence	Gene Promoter Element
Insulin	AGGCCC**TAAT**GGGCCA	A1
	AGACTC**TAAT**GACCCG	A3
	CTGGTC**TAAT**GTGGAA	A5
IAPP	GGAAAT**TAAT**GACAGA	A1
	ATGAGT**TAAT**GTAATA	A2
	CTACGT**TAAT**ATTTAC	A3
Glucokinase	CAGCTC**TAAT**GACAGG	UPE-3
GLUT2	AAGCAA**TAAT**AATAA	GLUT2TAAT
	CTC**TAAT**GAG	Consensus Sequence

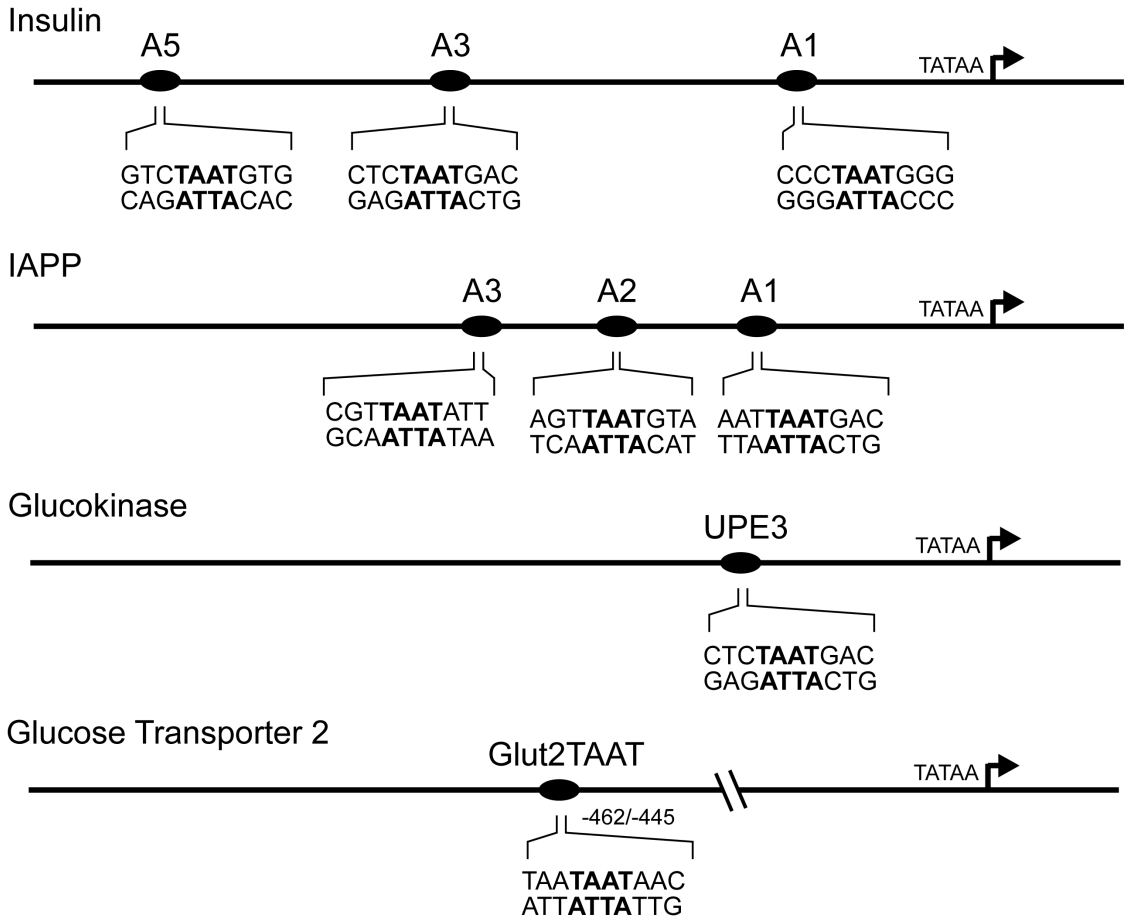

Figure **1-2.** Representation of the promoter region of the β-cell specific genes *insulin, islet amyloid polypeptide (IAPP), glucokinase, and glucose transporter 2.* Pdx1 binding elements are represented by black ovals and the sequence of each element is displayed below.

Early studies of Pdx1 were done by gel shift assays and cell based assays, in which Pdx1 was shown to interact with promoter binding sites containing the TAAT core and stimulate transcription of β-cell specific genes (Table 1-1). With the significant finding of Pdx1 expression in β-cells and its binding to AT-rich regions, sequences including the tetranucleotide TAAT sequence in a broad panel of β-cells specific gene promoters were investigated for binding by Pdx1. As with the insulin promoter studies, cell-based studies and *in vitro* binding assays were conducted to determine the interaction and effect of Pdx1. Many of these binding studies of Pdx1

were qualitative, but cognate binding sites were identified. Further functional studies of Pdx1 based on gel shift assays probed varying lengths of Pdx1, all encompassing the homeodomain motif. Overall, binding by Pdx1 to DNA sequences containing the core is fairly strong, within the nanomolar regime.[2, 27]

Knowing the promoter sites of various β-cell specific genes that Pdx1 is known to interact with in the context of the cell, Liberzon and colleagues completed a DNA sequence alignment of several of these binding sites (Table 1-2).[27] From the sequence alignment, several findings were presented. First, the consensus sequence of Pdx1 was qualitatively established as 5′-CTCTAATG(A/G)(C/G)-3′. *In vitro* binding studies of Pdx1 to several promoter elements and mutant DNA sequences were presented, in which the Pdx1 binding specificity to Ins A1 was notably tighter than any of the other sequences, including IAPP A2. More significantly, in the context of nucleotide sequence, Liberzon and colleagues presented a sequence alignment of several Pdx1 binding sites from promoters of different species, including rat.[27] Interestingly, the flanking DNA sequence in the 5′-region differed in rat and mouse from human in several of the binding sites (Table 1-2). It is important to note that many, if not most, of the studies of Pdx1 binding were in the context of the rat sequences, which not only differ from human in the 5′-region, but also an additional binding site adjacent to the A3 site is present (Table 1-2). This study provides speculation regarding how Pdx1 recognizes its cognate binding sequence and the manner in which Pdx1 interacts with the human promoter sequences. Thus, a significant part of this dissertation was made possible because of the interest in promoter sequence variation, in addition to identifying whether or not Pdx1 discriminates DNA sequences upstream from the core binding site and how.

Table **1-2.** Sequence alignment of A-box elements from different species. The alignment of Insulin and islet amyloid polypeptide (IAPP) promoters displays the deviations in flanking nucleotide sequence. [Modified from Liberzon, A. et al., 2004, NAR]

Insulin I	A1	rat	G C C C T **T A A T** G G G C C
		mouse	G A C C T **T A A T** G G G C C
		human	G G C C C **T A A T** G G G C C
	A3/	rat	T A A T C **T A A T** T A C C C
	A4	rat	G A C T C **T A A T** T A C C C
	A3/	mouse	G A C T A **T A A T** A A C C C
	A4	mouse	G A C T C **T A A T** T A C C C
	A3	human	G A C T C **T A A T** G A C C C
IAPP	A1	rat	G A C A C **T A A T** G A C A C
		mouse	G A C A C **T A A T** G A C A G
		human	G A A A T **T A A T** G A C A G
	A2	rat	A A G T G **T A A T** G T G T T
		mouse	A A G T G **T A A T** G T G T T
		human	T G A G T **T A A T** G T A A T

The structure of the homeodomain motif is well characterized; many, including Pdx1, have atomic resolution structures determined.[2] Although the unbound structure of Pdx1 has not been solved, the atomic resolution structure of the Pdx1-HD bound to a sequence containing the core binding site (5′-TAAT-3′) was reported (Figure 1-1B).[2] The X-ray structure of Pdx1 reveals the binding mode to a consensus sequence, in which sequence specific contacts with bases in the major groove are carried out by a series of conserved residues. In particular, Ile47, Gln50, Asn51, and Met54 interact via hydrogen bonds or van der Waals interactions with the DNA Watson-Crick face, resulting in direct readout of Ade2, Ade3, and Thy4.

As important as those residues are for A-box recognition, they are not the only essential interactions; the N-terminal arm of Pdx1 contains the strictly conserved arginine in position 5 required for binding. In the co-crystal structure of Pdx1, Arg5 interacts within the narrow minor groove of the DNA where it hydrogen bonds to the ribose sugar and O2 of Thy1, in addition to

the ribose sugar of Ade2. It is reported that an additional hydrogen bond with Arg5 and Gua-1 of the complementary strand is made; however, in the crystal structure model, this interaction is not clear. Because the structure reported based on crystallography is static, it is plausible Arg5 could make additional interactions not identified in the structure. Therefore, by solution NMR methods and isothermal titration calorimetric (ITC) methods, the role of Arg5 in minor groove interactions was further defined (see Chapter 2).

The homeodomain motif is an important and well-studied component, however, understanding it is insufficient to completely understand the structure-function relationships of these transcription factors. Pdx1 is 283 amino acids in length; the homeodomain being 60 residues is only ~20% of its primary sequence (Figure 1-4). The regions outside the homeodomain have not been structurally characterized, although bioinformatics analysis suggests that they exhibit disorder. Importantly, several biochemical studies have implicated these regions in protein-protein interactions.[28-34] For example, Pdx1 interacts with a protein subunit of a large ubiquitin ligase complex, in which the Pdx1 C-terminus is the substrate. In response to low glucose levels, an increase in the ubiquitination of Pdx1 is observed marking degradation of the protein, thereby regulating Pdx1 levels and modulating expression of key β-cell hormones.

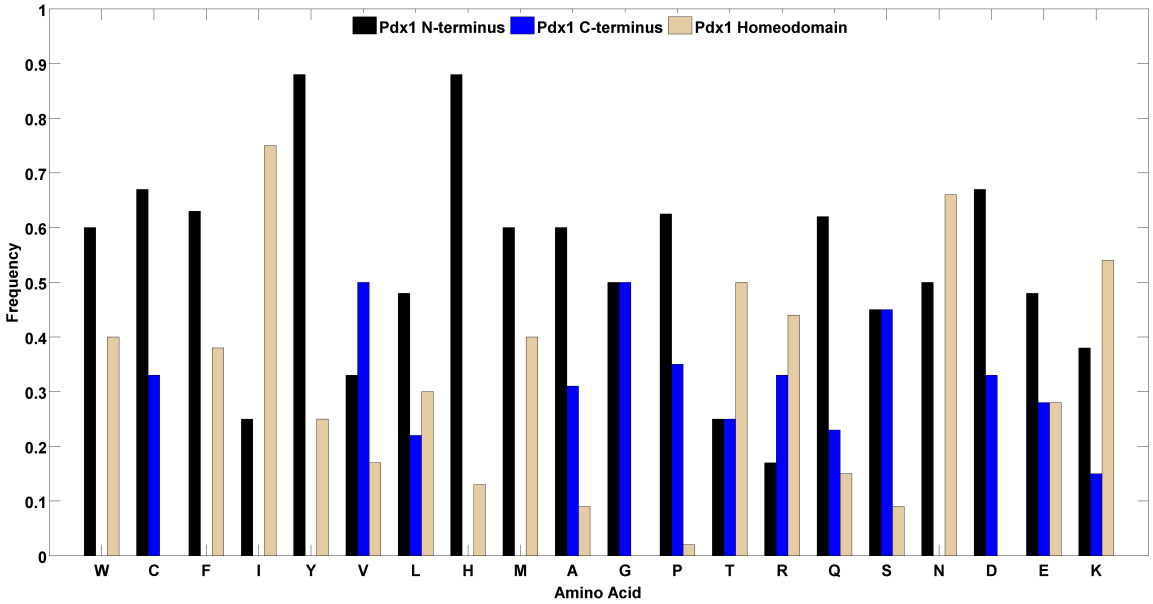

Figure **1-3.** Plot displaying the frequencies of a given amid acid represented within the three domains of Pdx1 − N-terminus (1-145), homeodomain (146-205), and C-terminus (206-283). Frequency was computed as the ratio of the occurrence of amino acid type in domain to total number of amino acid type in full length. The plot is segregated by hydrophobic residues on the left and polar residues on the right.

As mentioned in section 1.2, the amino acid sequence of intrinsically disordered regions shows a bias towards polar residues and in some cases, proline, glycines, and/or alanines. To visualize the amino acid frequencies of the three regions of Pdx1 (N-terminus, homeodomain, and C-terminus), a plot representing the ratio of the number of a specific amino acid in one of the 3 domains to the total number of the amino acid in the full length is shown in Figure 1-3. For example, ca. 35% of the proline residues in Pdx1 are located within Pdx1-C and ca. 60% of the prolines are in Pdx1-N. An inspection of Figure 1-3 shows sequence bias towards polar residues within Pdx1-C (i.e., depletion of hydrophobic residues), while most residues are similarly populated in Pdx1-N and Pdx1-HD. However, based on bioinformatics analysis, the structure predicted for Pdx1-N is dissimilar to the Pdx1-HD structure. Based on the disorder prediction algorithm PONDR, shown in Figure 1-4, the primary sequence of Pdx1 displays a substantial

amount of disorder evidenced by PONDR scores > 0.5. The baseline at 0.5, in Figure 1-4, represents the confidence score for disorder; values below 0.5 are predicted to have a tendency towards order, whereas values greater than 0.5 represent disorder. Therefore, the question that arises is: are these predicted disordered regions of Pdx1 truly disordered? Part of the work in this dissertation, relying heavily on nuclear magnetic resonance (NMR) techniques outlined below, provides evidence for structure of the C-terminal region.

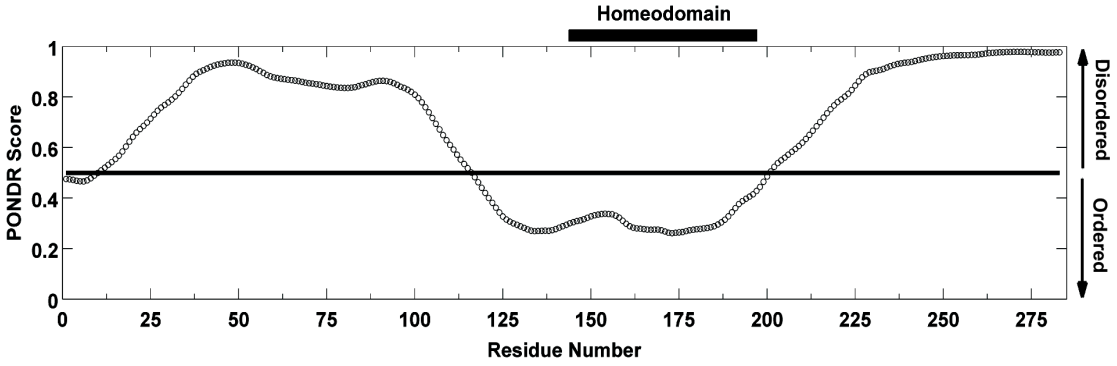

Figure **1-4.** Disorder Prediction of Pdx1 by PONDR. Values above a PONDR score of 0.5 indicate disordered residues and values below 0.5 indicate ordered residues.

1.3.1 Maturity-onset of the Young Type 4: A form of type II Diabetes

Mature onset diabetes of the young (MODY) is a form of type 2 diabetes that occurs in young children with mutations within transcription factors. Transcription factors bind specific DNA sequences with high affinity; mutant phenotypes result when protein functions are compromised, as with the R197H mutation located on the recognition motif of Pdx1 resulting in decreased activity.[35] Twelve mutations, which are illustrated in Figure 1-5, have been identified in the *pdx1* gene of diabetic patients; one of the mutations resulted in pancreatic agenesis followed by a diabetic phenotype.[35-41] Two of the 12 mutations are located in the homeodomain and result in decreased transactivation activity.[37, 42] The additional 10 are found in the disordered

regions not known to bind DNA. How these mutations affect binding and transactivation activity is not well understood. I hypothesize that these regions impart an organizational role by participating in protein-protein interactions, but their roles have not been sufficiently well defined to elucidate the phenotype. Therefore, in order to better understand how a mutation of Pdx1 leads to a MODY phenotype, a more complete structural and functional understanding of the disordered tails from Pdx1 must be developed. Following the studies presented in this dissertation, investigating the properties of MODY-associated Pdx1 mutants will shed light on the molecular origins of this disease.

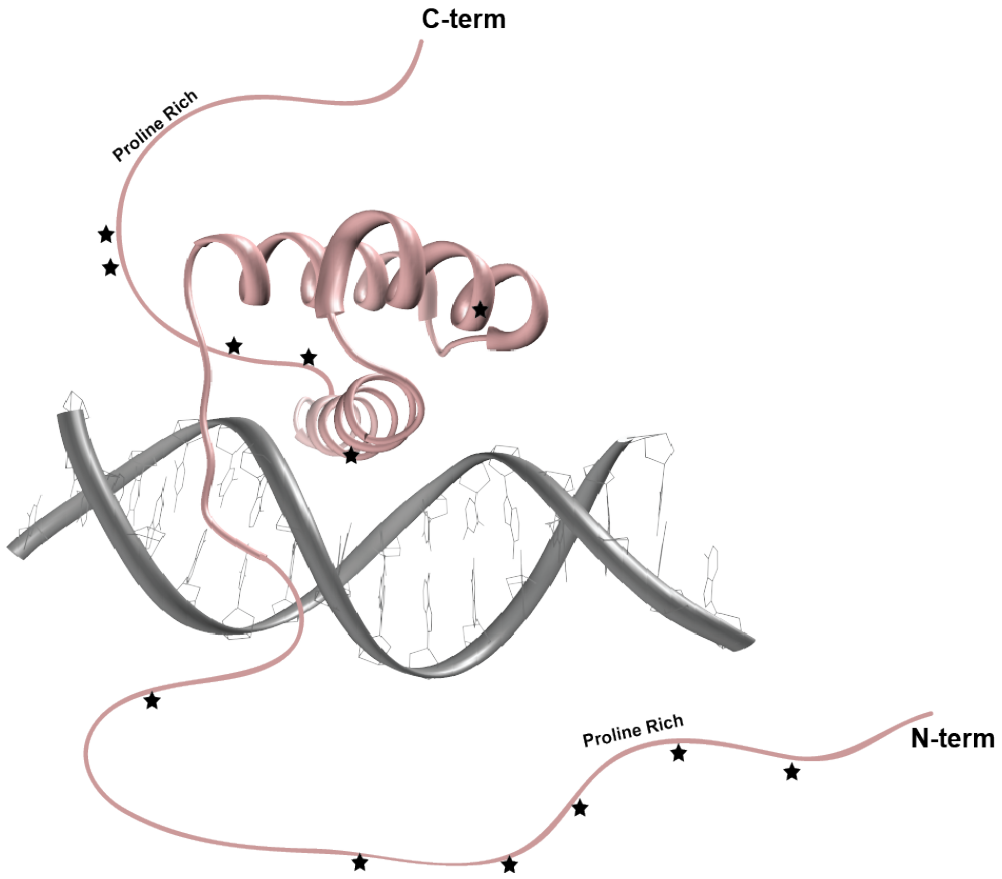

Figure **1-5.** Illustration of the structure of Pdx1bound to dsDNA. The disordered tails are represented by the long coil structures and the MODY4 mutations are marked by black stars. The core binding site of Pdx1 in indicated by the TAAT in the major groove. The N-terminus and C-terminus are proline-rich as indicated.

The goal of this dissertation was to be holistic in studying both the N-terminus and C-terminus. However, for the most part, my successes were focused on the C-terminus of Pdx1. The focus of this dissertation will be on the homeodomain and the C-terminus of Pdx1. First, however, the methodological framework for studying the structure, dynamics, and function of homeodomain transcription factors and Pdx1 is presented in the following section.

1.4 Biophysical Methods

1.4.1 Nuclear Magnetic Resonance Methods

Many investigators would argue that the function of well-folded proteins is easily deduced by their three dimensional conformation, but this is not so for IDPs. Classically, structures of cooperatively folded proteins were, and still are primarily, obtained by X-ray crystallographic methods. For crystallography to be of value, the protein or complex must be homogeneous to yield sharp diffraction patterns to 1-2 Å resolution for smaller systems. Structural heterogeneity or inherently flexible regions can prevent crystallization or will not yield appreciable electron density for structure generation. Despite these limitations, crystallography provides excellent structural details of biomolecules. However, the restrictions described above render X-ray crystallography unsuitable to study intrinsically disordered proteins.

NMR is a powerful and versatile method to study proteins that are, generally, < 25 KDa, especially those that do not form stable secondary or tertiary structures. Not only can NMR be used for structure determination, NMR can also be employed to study kinetics, dynamics, and binding, among others. Like with all techniques, there are some limitations to consider. NMR is a relatively insensitive technique, requiring high sample concentrations, which can be problematic

for some proteins. Typically, traditional proton-detection methods are employed to study well-folded proteins. However, proton-detection methods are not always optimal due to low signal dispersion and pH restrictions, as is the case with some intrinsically disordered proteins. Recent advancements in NMR pulse sequences featuring carbon-detection have made NMR a versatile tool to study IDPs.[43-47] The Showalter Laboratory has taken a leadership role in the development of these applications, driven in part by the work presented in chapters 4 and 5. Therefore, an introduction to our laboratory's carbon-detected NMR strategies will be described in the following sections.

1.4.1.1 Proton Based Nuclear Magnetic Resonance Techniques

Because proton methods typically yield great signal dispersion of individual amino acids for well-folded proteins and enhanced sensitivity, they are the most widely used methods for biological NMR. The $^{1}H,^{15}N$-HSQC for well-folded proteins spans a wide chemical shift range in the proton dimension (at least 6.5-9.5 ppm in most cases); however, the spectra of IDPs are much more compressed (chemical shifts within 7.5-8.5 ppm). For IDPs whose amino acid sequence is typical biased towards polar and/or charged residues, their $^{1}H,^{15}N$-HSQC results in spectral crowding, often with severe peak overlap.

Traditional proton NMR methods are routine for the sequential assignment of cooperatively folding proteins, such as the homeodomain of Pdx1.[48-52] Backbone assignments of proteins are completed prior to the collection of subsequent NMR experiments, such as spin dynamics (discussion in section 1.4.1.3.). For backbone assignment, a series of 2D and 3D experiments are collected. First, a 2D spectrum correlating the amide proton with the amide nitrogen ($^{1}H,^{15}N$-HSQC, in this case) of an individual amino acid is collected; this becomes the basis for subsequent experiments. The carbonyl carbon, Cα and Cβ chemical shifts are collected

to assign individual amino acids by the daisy chain method. Cα and Cβ chemical shifts aid in the identification of residue type. After identifying each amino acid in the 2D spectrum, data can be collected on a per residue basis.

Many homeodomain motifs from different proteins have been characterized by NMR. For example, the prototypical homeodomain Antp was structurally characterized by NMR, followed by a crystal structure.[4, 5, 53] The DNA-bound structures were structurally similar when compared. Because NMR can be used to study the structure, dynamics, and function of diverse protein conformations in a relatively efficient method, under favorable protein conditions this technique is ideal.

1.4.1.2. Carbon-Direct Detect Nuclear Magnetic Resonance Techniques

As was just reviewed, traditional proton detect experiments are typically not as useful for the backbone assignment or structure determination of IDPs owing to their degenerate sequence, which often leads to significant signal overlap in an $^{1}H,^{15}N$ spectrum. Furthermore, many IDPs are rich in proline residues, which are unobservable in a $^{1}H,^{15}N$ spectrum. This has provided the motivation for recent advances in NMR pulse sequences that have enabled the study of IDPs using carbon-detected NMR. These new experiments offer improved spectral dispersion and restore observation of proline residues by selection of the carbonyl carbon for detection in lieu of the amide proton (Figure 1-6). Moreover, structural and dynamics studies of IDPs are necessary to understand how the inherent plasticity of IDPs affects their function, how they interact with their partners, and ultimately their role in biological processes.

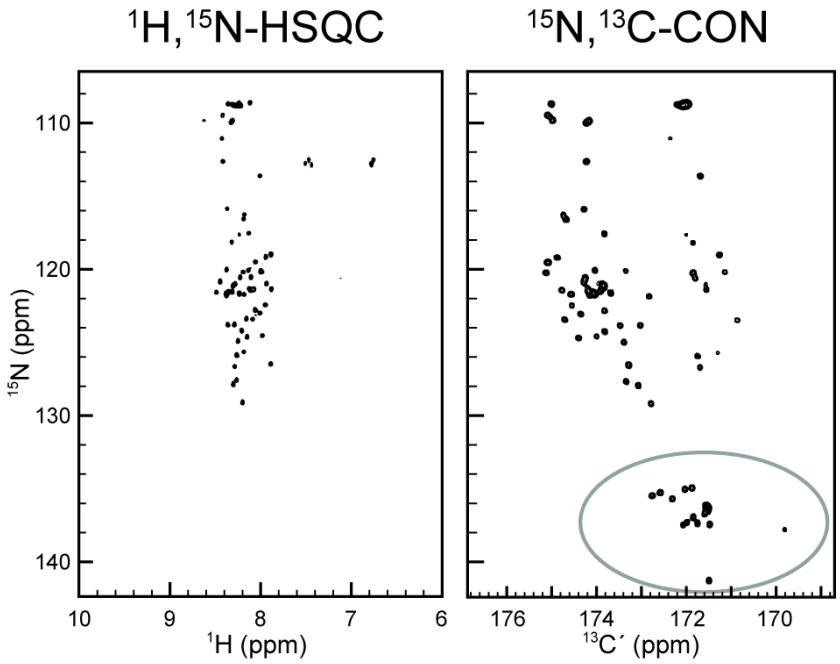

Figure **1-6.** Two-dimensional NMR spectra of the intrinsically disordered C-terminus of Pdx1. The traditional ^{1}H, ^{15}N-HSQC of Pdx1-C on the left and carbon-detect ^{15}N, ^{13}C-CON of Pdx1-C on the right. The proline residues are circled in the ^{15}N, ^{13}C-CON. Prolines are not observed in an ^{1}H, ^{15}N-HSQC due to the lack of the amide proton in the proline backbone.

Several recent studies of IDPs using carbon-detected NMR offer insight into the power that this method provides in studying disordered proteins.[46, 47, 54] The well-folded RAP74 domain binds to FCP1, a disordered protein. FCP1 samples many disordered conformations, but also samples an α-helix in the binding region. This was explored and validated by dynamic studies using carbon detected NMR methods. In particular, Lawrence and Showalter showed that the relaxation properties of residues in the binding region of FCP1 are different than residues in the non-binding region.[46] Structure prediction software calculates some degree of order in the binding region of FCP1, where the partial helix forms. Using carbon detect NMR experiments, the binding mechanism of the intrinsically disordered domain of hepatitis C NS5A was studied.[54] Part of this study focused on a proline rich region of NS5A, which was made possible by carbon detect methods.[47, 54] In particular, a polyproline region encompassing two prolines and

phosphorylated threonine were identified to participate in binding to a SH3 domain, which yielded a structural model.[54]

1.4.1.3. Spin Dynamics: Fast timescale dynamics of Proteins

NMR is a versatile method, which has been utilized to structurally and functionally characterize proteins. In a cellular environment, well-folded proteins are highly dynamic molecules; the dynamics of disordered proteins are even more extreme. Protein dynamics can be probed by a variety of methods, such as fluorescence methods (Förster resonance energy transfer (FRET), fluorescence anisotropy, time resolved fluorescence)[55-58], hydrogen/deuterium exchange mass spectrometry[59], electron paramagnetic resonance and NMR. A wide range of time scales (picoseconds to milliseconds) can be studied using NMR to investigate protein dynamics. The time scale of the dynamics studied by NMR informs us on individual amino acids and the global motions of a protein. Widespread use of the ^{15}N nuclei to study protein dynamics has been common since the first such study in 1989 by Kay and colleagues.[60] Moreover, using NMR to probe the dynamics of proteins, not only is the dynamics of individual amino acids monitored, but also the overall fluctuations of the protein can be calculated resulting in structural details.

Measurement of ^{15}N spin relaxation is a well-established technique used to monitor the conformational dynamics of proteins.[61-63] Protein backbone motions can be monitored over a wide range of time scales by measuring ^{15}N T_1, T_2, and ^{1}H-^{15}N NOE experiments.[60] Further analysis of ^{15}N relaxation data, by the Lipari-Szabo method yields information on the rigidity of protein backbones.[64] Unfortunately, it must be noted that the Lipari-Szabo model-free approach is only applicable if global and internal motions are rigorously separable; therefore, while useful, this ubiquitous literature method can only be applied to well-folded proteins.[65] Throughout this dissertation, ^{15}N is used as the probe to measure fast time scale dynamics of folded and unfolded

protein backbones to yield structural information, but at present no formalism to replace Lipari-Szabo is available for IDPs.

Altogether, spin relaxation of well-folded and disordered domains of proteins provides structural information without the need for structure determination. Moreover, it provides information about the inherent flexibility on a per residues basis that informs the structural role and often the functional role of certain residues. For the homeodomain, spin relaxation of the apo and DNA-bound state show a marked difference in the N-terminal arm that is responsible for DNA affinity and contributes to DNA specificity by interactions within the minor groove. Furthermore, spin relaxation can be just as powerful for disordered proteins as for well-folded proteins, especially when complementing the data with other methods. Overall, NMR methods provide a wide range of applications to study binding (apo vs holo), dynamics (spin relaxation), or structural information (spin relaxation and residual dipolar coupling) of biomolecules in solution.

1.4.2 Thermodynamics of Protein Interactions

Understanding the structure-function relationship of protein systems has provided an abundance of information regarding how proteins carry out functions, such as transcription, signaling, and replication. Often multi-molecule complexes must form in order to carry out these functions. These interactions are carried out by specific amino acids, shape, charge, or variations of these. The binding of two components, such as the homeodomain binding DNA or protein-protein interactions involving intrinsically disordered proteins can be characterized by several binding assays. Fluorescence assay[66, 67], filtering binding[68], gel shift assay[69-71], or ITC[72, 73] are some of the popular methods to measure binding of two molecules. While all these methods yield binding affinities, ITC is the only method whereby a complete collection of thermodynamic

parameters can be obtained in a single and expedient label-free experiment. Appendix A provides a more detailed account on ITC and tips and tricks for the user.

An ITC instrument has two identical cells that are connected by an electronic circuit, which either applies or decreases power to the sample cell to maintain the temperature difference at zero. In an ITC experiment, a ligand in the syringe is titrated into the macromolecule within the cell. Figure 1-7A represents an example of the protein (ligand) titrated into the dsDNA (macromolecule) generating heat, which is translated into the raw data shown in Figure 1-7B (left panel). During an experiment, the ligand is slowly titrated into the cell in a series of volumetric injections. The raw data is a compilation of averaged data points acquired. In an endothermic reaction, the sample cell temperature decreases due to the uptake of heat as a result of the reaction, thus an increase in power is necessary to maintain a change in temperature equivalent to zero between the two cells. For an exothermic reaction, the reaction generates heat, which increases the temperature of the sample cell, subsequently the power must decrease to cool the temperature within the sample cell to maintain a $\Delta T=0$. An example of an exothermic reaction is shown in Figure 1-7B, where the injections have negative values illustrating a decrease in power. The raw data in Figure 1-7B is integrated to obtain the heat of binding for each injection, resulting in a thermogram shown in the right panel of Figure 1-7B. As a result, the data is fit to various models, one-site, two-site, or sequential binding, for example, to attain the apparent binding enthalpy (ΔH), apparent binding association constant (K_a), and stoichiometry (n).

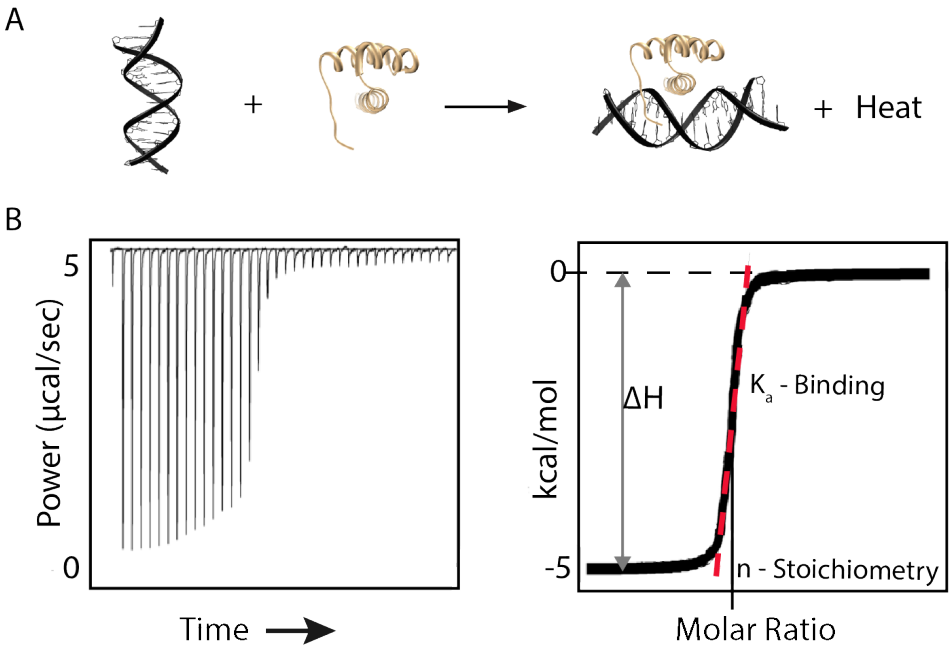

Figure **1-7.** Representation of an isothermal titration calorimetry experiment and data. (A) Representation of titration where ligand is injected into macromolecule and binding results in a release of heat. (B) Left panel displays the raw data for interaction in (A) and integration of raw data yields heat shown in the right panel. Thermodynamic parameters, ΔH, K_a, and n are derived from fitting the data to a binding model.

ITC has been utilized to probe the interaction thermodynamics of a variety of protein systems, from protein-small molecule interactions, nucleic acid binding proteins, protein-protein interactions of enzymes subunits to protein-protein interactions involving disordered proteins. The study of thermodynamics, by ITC, of DNA-binding proteins and their interactions with DNA and protein-protein interactions involving disordered proteins has been established. For most chapters within this dissertation, ITC was a main technique. Choosing ITC as the binding assay for this dissertation was wise, in that electrostatics played a major role in binding, especially for the homeodomain. The binding enthalpies became essential in the analyses involving the homeodomain.

Literature suggests that disordered tails of DNA-binding proteins may facilitate searches for specific sequences or modulate DNA binding.[74, 75] Pdx1 has long regions of disorder at both termini

of the homeodomain; however, their precise roles in Pdx1 function remain abstruse. Although the functional role of Pdx1's disordered regions have been implicated in protein-protein interactions, there is evidence that the disordered regions of homeodomain proteins play a regulatory role in DNA-binding. ITC will be employed to study the effects of the disordered C-terminus of Pdx1 on DNA binding and it's (assumed) binding to a protein implicated in the regulation of Pdx1 accumulation within β-cells. The protein-protein interaction with the C-terminus of Pdx1 may be partially electrostatic (as informed by literature) thus making calorimetry the most effective choice of method for the binding studies of the C-terminus of Pdx1.

1.5 Summary

The structure and function of Pdx1 is presented in this dissertation. The work described in Chapter 2 is an account of Pdx1's nucleotide preference studied by molecular dynamics calculations and thermodynamic experiments of Pdx1 binding with a panel of DNA sequences. In this chapter, we determined the preferential binding exhibited by Pdx1 on two different promoters and the effects of DNA and protein sequence on binding. The work presented in Chapter 3 is reports on the effects on DNA-binding by adding 5 residues to the C-terminus of the homeodomain from the disordered C-terminal domain of Pdx1. Protein stability played an important role in arriving at a more stable and inclusive construct. Again, a computational and thermodynamic approach was taken to arrive at the conclusion that regions outside the homeodomain tune DNA affinity. Chapter 4 provides a description of carbon-detected NMR studies of intrinsically disordered proteins, with the C-terminus of Pdx1 as an example. Chapter 5 is a structural and functional study of the intrinsically disordered C-terminus of Pdx1 based on NMR and ITC methods. Collectively, the data shows that this rather large region of the protein adopts a random coil structure that interacts with another protein. The future

directions for this project are described in Chapter 6. Finally, a description of ITC and best practices are listed in Appendix 1.

1.6 References

1. Gehring, W. J., Affolter, M., and Burglin, T. (1994) Homeodomain proteins, *Annual review of biochemistry 63*, 487-526.

2. Longo, A., Guanga, G. P., and Rose, R. B. (2007) Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain, *Biochemistry 46*, 2948-2957.

3. Mann, R. S. (1995) The specificity of homeotic gene function, *BioEssays : news and reviews in molecular, cellular and developmental biology 17*, 855-863.

4. Billeter, M., Qian, Y. Q., Otting, G., Muller, M., Gehring, W., and Wuthrich, K. (1993) Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex, *Journal of molecular biology 234*, 1084-1093.

5. Qian, Y. Q., Otting, G., Billeter, M., Muller, M., Gehring, W., and Wuthrich, K. (1993) Nuclear magnetic resonance spectroscopy of a DNA complex with the uniformly 13C-labeled Antennapedia homeodomain and structure determination of the DNA-bound homeodomain, *Journal of molecular biology 234*, 1070-1083.

6. Billeter, M., Qian, Y. Q., Otting, G., Muller, M., Gehring, W., and Wuthrich, K. (1993) Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex, *J Mol Biol 234*, 1084-1093.

7. White, S., Szewczyk, J. W., Turner, J. M., Baird, E. E., and Dervan, P. B. (1998) Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands, *Nature 391*, 468-471.

8. Gottesfeld, J. M., Neely, L., Trauger, J. W., Baird, E. E., and Dervan, P. B. (1997) Regulation of gene expression by small molecules, *Nature 387*, 202-205.

9. Dervan, P. B., and Edelson, B. S. (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides, *Current opinion in structural biology 13*, 284-299.

10. Dervan, P. B., and Burli, R. W. (1999) Sequence-specific DNA recognition by polyamides, *Current opinion in chemical biology 3*, 688-693.

11. Dervan, P. B. (2001) Molecular recognition of DNA by small molecules, *Bioorganic & medicinal chemistry 9*, 2215-2235.

12. Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009) The role of DNA shape in protein-DNA recognition, *Nature 461*, 1248-1253.

13. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function, *Biochemistry 41*, 6573-6582.

14. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes, *Genome Inform Ser Workshop Genome Inform 11*, 161-171.

15. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *Journal of molecular biology 293*, 321-331.

16. Sugase, K., Dyson, H. J., and Wright, P. E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein, *Nature 447*, 1021-1025.

17. Love, J. J., Li, X., Case, D. A., Giese, K., Grosschedl, R., and Wright, P. E. (1995) Structural basis for DNA bending by the architectural transcription factor LEF-1, *Nature 376*, 791-795.

18. Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C., and Asturias, F. J. (2008) Malleable machines take shape in eukaryotic transcriptional regulation, *Nature chemical biology 4*, 728-737.

19. Forman-Kay, J. D., and Mittag, T. (2013) From sequence and forces to structure, function, and evolution of intrinsically disordered proteins, *Structure 21*, 1492-1499.

20. Gibbs, E. B., and Showalter, S. A. (2015) Quantitative biophysical characterization of intrinsically disordered proteins, *Biochemistry 54*, 1314-1326.

21. Miller, C. P., McGehee, R. E., Jr., and Habener, J. F. (1994) IDX-1: a new homeodomain transcription factor expressed in rat pancreatic islets and duodenum that transactivates the somatostatin gene, *The EMBO journal 13*, 1145-1156.

22. Babu, D. A., Deering, T. G., and Mirmira, R. G. (2007) A feat of metabolic proportions: Pdx1 orchestrates islet development and function in the maintenance of glucose homeostasis, *Molecular genetics and metabolism 92*, 43-55.

23. Watada, H., Kajimoto, Y., Miyagawa, J., Hanafusa, T., Hamaguchi, K., Matsuoka, T., Yamamoto, K., Matsuzawa, Y., Kawamori, R., and Yamasaki, Y. (1996) PDX-1 induces insulin and glucokinase gene expressions in alphaTC1 clone 6 cells in the presence of betacellulin, *Diabetes 45*, 1826-1831.

24. Watada, H., Kajimoto, Y., Umayahara, Y., Matsuoka, T., Kaneto, H., Fujitani, Y., Kamada, T., Kawamori, R., and Yamasaki, Y. (1996) The human glucokinase gene beta-cell-type promoter: an essential role of insulin promoter factor 1/PDX-1 in its activation in HIT-T15 cells, *Diabetes 45*, 1478-1488.

25. Watada, H., Kajimoto, Y., Kaneto, H., Matsuoka, T., Fujitani, Y., Miyazaki, J., and Yamasaki, Y. (1996) Involvement of the homeodomain-containing transcription factor PDX-1 in islet amyloid polypeptide gene transcription, *Biochemical and biophysical research communications 229*, 746-751.

26. Waeber, G., Thompson, N., Nicod, P., and Bonny, C. (1996) Transcriptional activation of the GLUT2 gene by the IPF-1/STF-1/IDX-1 homeobox factor, *Molecular endocrinology 10*, 1327-1334.

27. Liberzon, A., Ridner, G., and Walker, M. D. (2004) Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1, *Nucleic acids research 32*, 54-64.

28. Liu, A., Desai, B. M., and Stoffers, D. A. (2004) Identification of PCIF1, a POZ domain protein that inhibits PDX-1 (MODY4) transcriptional activity, *Molecular and cellular biology 24*, 4372-4383.

29. Liu, A., Oliver-Krasinski, J., and Stoffers, D. A. (2006) Two conserved domains in PCIF1 mediate interaction with pancreatic transcription factor PDX-1, *FEBS letters 580*, 6701-6706.

30. Qiu, Y., Guo, M., Huang, S., and Stein, R. (2002) Insulin gene transcription is mediated by interactions between the p300 coactivator and PDX-1, BETA2, and E47, *Molecular and cellular biology 22*, 412-420.

31. Stanojevic, V., Yao, K. M., and Thomas, M. K. (2005) The coactivator Bridge-1 increases transcriptional activation by pancreas duodenum homeobox-1 (PDX-1), *Molecular and cellular endocrinology 237*, 67-74.

32. Mosley, A. L., and Ozcan, S. (2004) The pancreatic duodenal homeobox-1 protein (Pdx-1) interacts with histone deacetylases Hdac-1 and Hdac-2 on low levels of glucose, *The Journal of biological chemistry 279*, 54241-54247.

33. Mosley, A. L., Corbett, J. A., and Ozcan, S. (2004) Glucose regulation of insulin gene expression requires the recruitment of p300 by the beta-cell-specific transcription factor Pdx-1, *Molecular endocrinology 18*, 2279-2290.

34. Peshavaria, M., Henderson, E., Sharma, A., Wright, C. V., and Stein, R. (1997) Functional characterization of the transactivation properties of the PDX-1 homeodomain protein, *Molecular and cellular biology 17*, 3987-3996.

35. Hansen, L., Urioste, S., Petersen, H. V., Jensen, J. N., Eiberg, H., Barbetti, F., Serup, P., Hansen, T., and Pedersen, O. (2000) Missense mutations in the human insulin promoter factor-1 gene and their relation to maturity-onset diabetes of the young and late-onset type 2 diabetes mellitus in caucasians, *J Clin Endocrinol Metab 85*, 1323-1326.

36. Al-Quobaili, F., and Montenarh, M. (2008) Pancreatic duodenal homeobox factor-1 and diabetes mellitus type 2, *International Journal of Molecular Medicine 2008*, 399.

37. Hani, E. H., Stoffers, D. A., Chevre, J. C., Durand, E., Stanojevic, V., Dina, C., Habener, J. F., and Froguel, P. (1999) Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus, *J Clin Invest 104*, R41-48.

38. Macfarlane, W. M., Frayling, T. M., Ellard, S., Evans, J. C., Allen, L. I., Bulman, M. P., Ayres, S., Shepherd, M., Clark, P., Millward, A., Demaine, A., Wilkin, T., Docherty, K., and Hattersley, A. T. (1999) Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes, *J Clin Invest 104*, R33-39.

39. Stoffers, D. A., Ferrer, J., Clarke, W. L., and Habener, J. F. (1997) Early-onset type-II diabetes mellitus (MODY4) linked to IPF1, *Nat Genet 17*, 138-139.

40. Weng, J., Macfarlane, W. M., Lehto, M., Gu, H. F., Shepherd, L. M., Ivarsson, S. A., Wibell, L., Smith, T., and Groop, L. C. (2001) Functional consequences of mutations in the MODY4 gene (IPF1) and coexistence with MODY3 mutations, *Diabetologia 44*, 249-258.

41. Gragnoli, C., Lindner, T., Chiaromonte, F., Colasurdo, L., Dantonio, T., Gragnoli, F., Gragnoli, G., Manetti, E., Signorini, A. M., and Marozzi, G. (1998) Identification of a missense mutation (P33T) in the insulin promoter factor-1 (IPF-1) gene in an Italian patient with early onset type 2 diabetes, *Diabetologia 41*, 412-412.

42. Nicolino, M., Claiborn, K. C., Senee, V., Boland, A., Stoffers, D. A., and Julier, C. (2010) A novel hypomorphic PDX1 mutation responsible for permanent neonatal diabetes with subclinical exocrine deficiency, *Diabetes 59*, 733-740.

43. Bermel, W., Bertini, I., Felli, I. C., Lee, Y. M., Luchinat, C., and Pierattelli, R. (2006) Protonless NMR experiments for sequence-specific assignment of backbone nuclei in unfolded proteins, *Journal of the American Chemical Society 128*, 3918-3919.

44. Bermel, W., Bertini, I., Chill, J., Felli, I. C., Haba, N., Kumar, M. V. V., and Pierattelli, R. (2012) Exclusively heteronuclear (13) C-detected amino-acid-selective NMR experiments for the study of intrinsically disordered proteins (IDPs), *Chembiochem : a European journal of chemical biology 13*, 2425-2432.

45. Bermel, W., Bertini, I., Felli, I. C., and Pierattelli, R. (2009) Speeding up (13)C direct detection biomolecular NMR spectroscopy, *Journal of the American Chemical Society 131*, 15339-15345.

46. Lawrence, C. W., and Showalter, S. A. (2012) Carbon-Detected 15N NMR Spin Relaxation of an Intrinsically Disordered Protein: FCP1 Dynamics Unbound and in Complex with RAP74, *Journal of Physical Chemistry Letters 3*, 1409-1413.

47. Sahu, D., Bastidas, M., and Showalter, S. A. (2014) Generating NMR chemical shift assignments of intrinsically disordered proteins using carbon-detected NMR methods, *Analytical biochemistry 449*, 17-25.

48. Kay, L. E., Ikura, M., Tschudin, R., and Bax, A. (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins, *Journal of Magnetic Resonance 89*, 496-514.

49. Grzesiek, S., and Bax, A. (1992) Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein, *Journal of Magnetic Resonance 96*, 432-440.

50. Clubb, R. T., Thanabal, V., and Wagner, G. (1992) A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN and 13C chemical shifts in 15N-13C-labeled proteins, *journal of magnetic resonance 97*, 213-217.

51. Grzesiek, S., and Bax, A. (1992) Correlating backbone amide and side chain resoances in larger proteins by multiple relayed triple resonace NMR, *Journal of the American Chemical Society 114*, 6291-6293.

52. Grzesiek, S., and Bax, A. (1992) An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins, *J Magn Reson 99*, 201-207.

53. Fraenkel, E., and Pabo, C. O. (1998) Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex, *Nature structural biology 5*, 692-697.

54. Cordek, D. G., Croom-Perez, T. J., Hwang, J., Hargittai, M. R., Subba-Reddy, C. V., Han, Q., Lodeiro, M. F., Ning, G., McCrory, T. S., Arnold, J. J., Koc, H., Lindenbach, B. D., Showalter, S. A., and Cameron, C. E. (2014) Expanding the proteome of an RNA virus by phosphorylation of an intrinsically disordered viral protein, *The Journal of biological chemistry 289*, 24397-24416.

55. Heyduk, T. (2002) Measuring protein conformational changes by FRET/LRET, *Current opinion in biotechnology 13*, 292-296.

56. MacKerell, A. D., Jr., Rigler, R., Nilsson, L., Hahn, U., and Saenger, W. (1987) Protein dynamics. A time-resolved fluorescence, energetic and molecular dynamics study of ribonuclease T1, *Biophysical chemistry 26*, 247-261.

57. Schroder, G. F., Alexiev, U., and Grubmuller, H. (2005) Simulation of fluorescence anisotropy experiments: probing protein dynamics, *Biophysical journal 89*, 3757-3770.

58. Lakowicz, J. R., Cherek, H., Gryczynski, I., Joshi, N., and Johnson, M. L. (1987) Enhanced resolution of fluorescence anisotropy decays by simultaneous analysis of progressively quenched samples. Applications to anisotropic rotations and to protein dynamics, *Biophysical journal 51*, 755-768.

59. Tsutsui, Y., and Wintrode, P. L. (2007) Hydrogen/deuterium exchange-mass spectrometry: a powerful tool for probing protein structure, dynamics and interactions, *Current medicinal chemistry 14*, 2344-2358.

60. Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease, *Biochemistry 28*, 8972-8979.

61. Palmer, A. G., 3rd. (2004) NMR characterization of the dynamics of biomacromolecules, *Chemical reviews 104*, 3623-3640.

62. Jarymowycz, V. A., and Stone, M. J. (2006) Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences, *Chemical reviews 106*, 1624-1671.

63. Peng, J. W. (2012) Exposing the Moving Parts of Proteins with NMR Spectroscopy, *The journal of physical chemistry letters 3*, 1039-1051.

64. Lipari, G., and Szabo, A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic- Resonance Relaxation in Macromolecules .1. Theory and Range of Validity, *Journal of the American Chemical Society 104*, 4546-4559.

65. Lipari, G., and Szabo, A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules .1. Theory and Range of Validity, *J Am Chem Soc 104*, 4546-4559.

66. Heyduk, T., Ma, Y., Tang, H., and Ebright, R. H. (1996) Fluorescence anisotropy: rapid, quantitative assay for protein-DNA and protein-protein interaction, *Methods in enzymology 274*, 492-503.

67. LeTilly, V., and Royer, C. A. (1993) Fluorescence anisotropy assays implicate protein-protein interactions in regulating trp repressor DNA binding, *Biochemistry 32*, 7753-7758.

68. Wong, I., and Lohman, T. M. (1993) A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions, *Proceedings of the National Academy of Sciences of the United States of America 90*, 5428-5432.

69. Wostenberg, C., Lary, J. W., Sahu, D., Acevedo, R., Quarles, K. A., Cole, J. L., and Showalter, S. A. (2012) The role of human Dicer-dsRBD in processing small regulatory RNAs, *PloS one 7*, e51829.

70. Quarles, K. A., Sahu, D., Havens, M. A., Forsyth, E. R., Wostenberg, C., Hastings, M. L., and Showalter, S. A. (2013) Ensemble analysis of primary microRNA structure reveals an extensive capacity to deform near the Drosha cleavage site, *Biochemistry 52*, 795-807.

71. Buratowski, S., and Chodosh, L. A. (2001) Mobility shift DNA-binding assay using gel electrophoresis, *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] Chapter 12*, Unit 12 12.

72. Buurma, N. J., and Haq, I. (2007) Advances in the analysis of isothermal titration calorimetry data for ligand-DNA interactions, *Methods 42*, 162-172.

73. Bastidas, M., and Showalter, S. A. (2013) Thermodynamic and structural determinants of differential Pdx1 binding to elements from the insulin and IAPP promoters, *Journal of molecular biology 425*, 3360-3377.

74. Liberzon, A., Ridner, G., and Walker, M. D. (2004) Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1, *Nucleic Acids Res 32*, 54-64.

75. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm, *J Mol Biol 293*, 321-331.

Chapter 2

Thermodynamic and Structural Determinants of Differential Pdx1 Binding to Elements from the Insulin and IAPP Promoters

[Published as a paper titled "Thermodynamics and Structural Determinants of Differential Pdx1 Binding to Elements from the Insulin and IAPP Promoters" by **Monique Bastidas**, and Scott A. Showalter in *J. Mol. Biol.*, (2013) 425, 3360-3377.] Molecular dynamics simulations carried out by Scott A. Showalter.

Abstract

In adult mammals, the production of insulin and other peptide hormones, such as the islet amyloid polypeptide (IAPP), is limited to β-cells due to tissue-specific expression of a set of transcription factors, the best known of which is Pdx1. Like many homeodomain transcription factors, Pdx1 binds to a core DNA recognition sequence containing the tetranucleotide 5′-TAAT-3′; its consensus recognition element is 5′-CTCTAAT(T/G)AG-3′. Currently, a complete thermodynamic profile of Pdx1 binding to near-consensus and native promoter sequences has not been established, obscuring the mechanism of target site selection by this critical transcription factor. Strikingly, while Pdx1 responsive elements in the human *insulin* promoter conform to the pentanucleotide 5′-CTAAT-3′ sequence, the Pdx1 responsive elements in the human *iapp* promoter all contain a substitution to 5′-TTAAT-3′. The crystal structure of Pdx1 bound to the consensus nucleotide sequence does not explain how Pdx1 identifies this natural variation, if it does at all. Here we report a combination of isothermal calorimetric titrations, NMR spectroscopy, and extensive multi-microsecond molecular dynamics calculations of Pdx1 that define its interactions with a panel of natural promoter elements and consensus-derived sequences. Our results show a small preference of Pdx1 for a Cyt base 5′

relative to the core TAAT promoter element. Molecular mechanics calculations, corroborated by experimental NMR data, lead to a rational explanation for sequence discrimination at this position. Taken together, our results suggest a molecular mechanism for differential Pdx1 affinity to elements from the *insulin* and *iapp* promoter sequences.

2.1 Introduction

Diabetes mellitus is a complex metabolic disease characterized by persistent hyperglycaemia resulting from either insufficient quantities of insulin, deficient insulin response in peripheral tissue, or both. In adult mammals, insulin production is limited to β-cells due to restricted expression of a set of tissue-specific transcription factors.[1] While the primary endocrine function of the β-cells is insulin secretion in response to increases in blood glucose levels, these crucial cells also release small amounts of other peptide hormones, including the islet amyloid polypeptide (IAPP, or amylin).[2, 3] Due to their co-storage in β-cell secretory granules, the concentrations of insulin and IAPP in the blood change in parallel in response to glucose stimulation,[4] supporting the hypothesis that IAPP is involved in the regulation of glucose metabolism.[5] IAPP is a regulatory peptide that functions in energy homeostasis, primarily by signaling satiation.[6] Binding of IAPP in the brain inhibits gastric emptying and elicits an anorectic response, resulting in reduced food intake.[7,8] Gastric emptying is pathologically rapid in type 1 diabetes, suggesting a phenotypic role for the absence of IAPP in this form of the disease.[9] Strikingly, both the *iapp* and *insulin* genes contain many similar promoter elements that regulate the effects of glucose on their transcription.[10, 11]

While no single protein or multi-protein complex alone accounts for cell-specific insulin or IAPP expression, the pancreatic duodenal homeobox protein 1 (Pdx1; also known as Iuf1, Ipf1, Idx1, Stf1, and Gsf) fulfills a central role in pancreatic development, endocrine pancreas maintenance, and activation of the *insulin* and *iapp* genes.[12,13,14,15] Further, it is known that the Pdx1 DNA-binding

activity directly induces both insulin and IAPP production in response to glucose.[16; 17] The consensus Pdx1 binding site has previously been reported as 5′-CTCTAAT(T/G)AG-3′.[18; 19] Natural sequences similar to this consensus are found in both the *insulin* and *iapp* gene promoters (Figure 2-1). Variation in the TAAT "core" recognition element, which is a required motif bound by most homeobox transcription factors, is poorly tolerated; variation in the peripheral positions of this core sequence has been documented to produce ten-fold increases in measured *in vitro* Pdx1 dissociation constants (K_d) that manifest as substantial reductions in promoter activity *in vivo*.[19] Once recruited to a promoter through its homeodomain, Pdx1 also interacts with multiple additional regulatory proteins, allowing synergistic interactions and modification of transcription outcomes in response to the current environment.[20]

Given that Pdx1 is crucial for the expression of genes central to the mature β-cell phenotype, it is surprising that many of the details of the molecular mechanism for Pdx1 influence on gene expression remain to be defined.[21] In this regard, the recently published crystal structure of Syrian hamster Pdx1 bound to a DNA duplex containing the consensus Pdx1 binding site provided a major breakthrough, by suggesting the atomic origins for much of the sequence specificity imparted by elements flanking the TAAT core.[22] To briefly summarize, Longo *et al.*[22] report that all three nucleobases in the strand opposite the consensus GAG sequence (found on the 3′-side relative to the TAAT motif) are directly contacted by residues Gln-195 and Met-199 of the homeodomain.[a] Further, Longo *et al.* suggest that the C·G base pair immediately 5′ to the TAAT must be recognized by the side chain of Arg-150, which is found in the intrinsically disordered N-terminal arm of the homeodomain (see Figure 2-1). This assertion was made despite an absence of direct contacts between Arg-150 and either nucleobase of the C·G pair in the deposited structure model.[22] Strikingly, the Pdx1 responsive elements in the promoter of the human *iapp* gene all contain a substitution from

[a] Throughout this paper we use numbering that indicates a residue's position within the human Pdx1 sequence. Residue 1 of the canonical 60 residue homeodomain sequence corresponds to residue number 146 of Pdx1.

5′-CTAAT-3′ to 5′-TTAAT-3′, underscoring the need to more fully understand nucleotide specificity in this position.

Currently, there is no rigorous molecular mechanism defining and rationalizing variation in Pdx1 affinity for DNA sequences similar to the consensus motif, beyond the preliminary information provided by elucidation of the Pdx1-homeodomain structure in complex with consensus DNA.[22] Thus, a systematic study of interactions involving non-consensus DNA sequences is needed to provide a rationale for understanding differential affinity of Pdx1 to disparate promoters that it regulates. Importantly, while the *insulin* promoter and the factors that bind it are well studied, far less effort has been put into fully characterizing the regulation of promoters for other proteins with significant roles in maintaining β-cell function, including the *iapp* promoter.[12] Functional assays illustrating differential Pdx1 stimulated transcription levels have been reported that anecdotally correlate with differences in AT-rich sequence,[23; 24; 25] but the results have not been directly verified through *in vitro* biophysical methods.

The present study aims to provide a molecularly detailed characterization of the differences between Pdx1 binding to regulated promoter elements from the *insulin* and *iapp* genes. Here we report the thermodynamics of Pdx1 binding to both consensus and human promoter derived DNA sequences, assayed by isothermal titration calorimetry (ITC). Our results show a preference of Pdx1 for a Cyt in the 5′-position with respect to the core TAAT promoter element. Additionally, we report NMR spin relaxation measurements and extensive multi-microsecond molecular dynamics calculations of Pdx1, both in the unbound state and bound to consensus DNA, which lead to a mechanistic hypothesis for the origins of sequence discrimination 5′ to the TAAT element. Our results rationalize the role of the intrinsically disordered N-terminal arm of Pdx1 in providing differential affinity for sites from the *insulin* and *iapp* gene promoters.

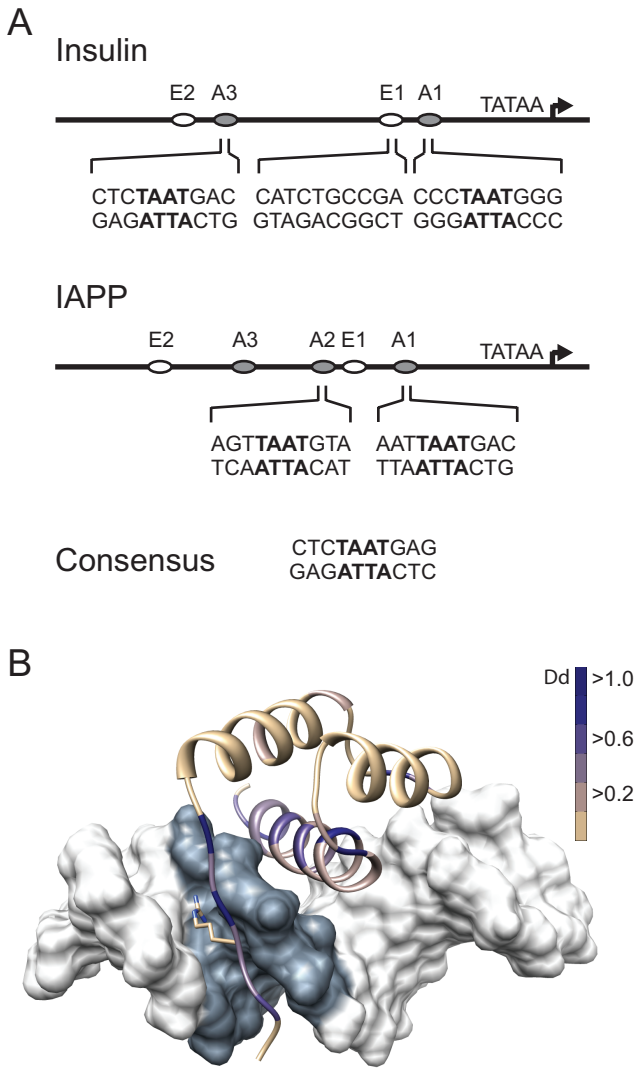

Figure **2-1.** Pdx1 recognizes A-box motifs common to the *insulin* and *iapp* gene promoters. (A) Schematic representations of the *insulin* and *iapp* gene promoters depict the relative location of A-box promoter elements, and the adjacent E-box elements, which are also involved in regulating transcription from these promoters. The nucleotide sequence in the neighborhood of the core TAAT element (denoted in bold) is shown for each of the four A-boxes investigated in this study. Pdx1 does not bind to E-boxes and so the *insulin* E1 sequence is used as a control for non-specific DNA binding in this study. (B) The DNA binding epitope of Pdx1 is confirmed by monitoring changes in the backbone amide chemical shift between the unbound and consensus-DNA bound states. Chemical shift changes are mapped as progression from tan to blue onto a ribbon-representation of Pdx1 in complex with duplex-DNA (molecular surface of pdb 2h1k, in which the core TAAT element is depicted in grey). Also depicted is the side chain of Arg-150, which is shown in the crystallographic orientation.

2.2 Results and Discussion

2.2.1 NMR verification of the solution binding interface

The crystal structure of Syrian hamster Pdx1 bound to a DNA duplex containing the consensus Pdx1 binding site adopts the canonical homeodomain fold, with three segments of α-helix and an extended N-terminal tail that inserts into the minor groove of the DNA.[22] The crystal structure of Pdx1 homeodomain in complex with a consensus DNA sequence (referred to here as DNA_{CON}) revealed that Pdx1 adopts the canonical homeodomain binding mode, with helix-3 inserting into the major groove of the DNA duplex and the intrinsically disordered N-terminal arm of the domain, which bears the strongly conserved residue Arg-150, bound to the minor groove (Figure 2-1B). We have verified that the co-crystal structure accurately represents the solution-state of the Pdx1-DNA_{CON} complex by solution NMR spectroscopy. Standard double and triple resonance spectra of apo-Pdx1 and the Pdx1-DNA_{CON} complex were acquired on an 850 MHz spectrometer, yielding complete backbone assignments of both states (although we note that in apo-Pdx1 Tyr-153 is the closest residue to the N-terminus yielding quantitative results at our experimental pH, due to rapid solvent exchange in this highly exposed region). DNA binding by Pdx1 resulted in a specific and temporally stable complex, as evidenced by the sharp 2D-lineshapes in the $^{1}H,^{15}N$-HSQC of the complex and by the change in peak distribution relative to the unbound state. These results are in agreement with the DNA-bound crystal structure of Pdx1 and with our ITC results (*vide infra*). Chemical shift perturbations are mapped onto the structure of Pdx1, represented as the complex with dsDNA in Figure 2-1B. Almost all residues displaying large changes in backbone amide chemical shift interact with the core TAAT sequence element of the DNA (shaded dark grey in Figure 2-1B), or are adjacent to those interaction sites, and are located in helix-3 or the N-terminal arm.

2.2.2 Thermodynamic analysis of promoter-derived DNA binding by Pdx1

Pdx1 binds to multiple sites within the promoters of β-cell specific genes, including *insulin* and *iapp*, where the bound DNA elements show sequence variation at positions believed to be recognized with nucleotide-specificity by Pdx1. Thus, characterization of binding to an idealized consensus sequence alone is not sufficient to describe the biological role of this factor well; it is important to characterize the thermodynamics of Pdx1 interactions with a set of related DNA sequences derived from natural promoter elements, as well as its interaction with sequences perturbed by point mutation.

Sequences similar to the consensus 5´-CTCTAAT(T/G)AG-3´ recognition element, generally referred to as A-boxes, are commonly found in multiple copies throughout the proximal promoter regions of β-cell associated genes. For example, A-boxes are found within the -400 to +1 region of the *insulin* and *iapp* gene promoters,[11; 26] the relative positions and exact sequences of which are provided in Figure 2-1. Multiple groups have reported equilibrium dissociation constants for Pdx1 binding to the A1 and A3 boxes from the rat and mouse *insulin* promoters, finding through competition gel mobility shift assays that the affinity is typically on the order of 1-5 nM.[19; 22; 27; 28] In contrast, affinities of Pdx1 for neighboring E-box elements, lacking the core TAAT element, have been reported to range from 20-150 nM by similar assays,[19; 22] suggesting that the *in vitro* specificity of Pdx1 for the A-box is not dramatic.

Table **2-1.** Thermodynamic parameters describing binding of Pdx1-HD to dsDNA derived from natural human promoter elements; measured at 298K in 100 mM cacodylate, pH 7.3, 100 mM KCl. The core TAAT recognition element is shown in bold (except in the E1 element, which lacks it, and is used in this study as a control to describe non-specific binding). All error bars on fitted ITC parameters represent standard error of the mean estimated from triplicate measurements, with fitting performed in Origin 7.0 (MicroCal, Inc.).

DNA	**Forward Sequence**	**n**	**K_d (nM)**	**ΔH (kcal mol^{-1})**
Insulin A1	5′-AGGCCC**TAAT**GGGCCA-3′	0.9 ± 0.1	7 ± 2	-10.2 ± 0.1
Insulin A3	5′-AGACTC**TAAT**GACCCG-3′	0.9 ± 0.1	4 ± 1	-9.63 ± 0.05
Insulin E1	5′-AGCCATCTGCCGACCC-3′	1.0 ± 0.1	1300 ± 300	-2.7 ± 0.1
IAPP A1	5′-GGAAAT**TAAT**GACAGA-3′	1.1 ± 0.1	22 ± 5	-4.60 ± 0.03
IAPP A2	5′-ATGAGT**TAAT**GTAATA-3′	1.2 ± 0.1	27 ± 7	-5.71 ± 0.05

Significantly, the A-box elements of the *insulin* and *iapp* gene promoters diverge at the 5′-position relative to the core TAAT recognition element (Figure 2-1), underscoring the importance of determining whether Pdx1 is sensitive to the nucleotide identity of this region. The co-crystal structure of Pdx1 bound to consensus DNA offers little guidance on this key issue. The crystallographic orientation of the side chain of Arg-150 facilitates engagement in hydrogen bonding with the Thy base of the 5′-base pair of the TAAT element exclusively; neither Arg-150 nor any other Pdx1 residue is in direct contact with the C·G base pair to the 5′-side of the core TAAT element. Here we have performed isothermal titration calorimetry (ITC) experiments at 298K to measure the interaction of Pdx1 with a total of 16 DNA sequences derived both from natural human promoter elements (Table 2-1) and from systematic point-mutation of the consensus sequence (Table 2-2). For each interaction, fitting parameters proportional to the binding enthalpy (ΔH), binding affinity (as reported by the equilibrium dissociation constant, K_d), and stoichiometry of interaction (n) are reported. The determined parameters reveal patterns in nucleotide sequence recognition that rationalize previously estimated levels of *in vivo* gene activation by Pdx1.

The overall trends in our data set are summarized by the representative titrations presented in Figure 2-2 for Pdx1 binding to Insulin-A3, IAPP-A1, the DNA_{CON} sequence, and Insulin-E1. A comprehensive list of best fit parameters for the set of titrations involving natural promoter sequences is provided in Table 2-1. Among the natural A-box sequences investigated here, Insulin-A3 and IAPP-A1 bear the most sequence identity to one another and with respect to the consensus high-affinity sequence: although their 5´-sequences diverge, both elements contain the required TAAT core and are identical through the first three base pairs on the 3´-side of the core element; the first two of these shared 3´-nucleotides are also identical in sequence to the consensus. Insulin-A3 and DNA_{CON} are nearly identical in the Pdx1 binding region, making it unsurprising that the measured dissociation constants for Pdx1 binding to both were indistinguishable within experimental uncertainty. In contrast, we find an approximate five-fold lower affinity of Pdx1 for IAPP-A1 than for Insulin-A3, or for DNA_{CON}, suggesting that one or more nucleotides on the 5´-side of the core TAAT element contribute significantly to sequence recognition by Pdx1. This result generalizes in that both *insulin* derived A-boxes studied contain the core pentanucleotide 5´-CTAAT-3´ sequence and were found to bind with affinities indistinguishable within error from that of the consensus element, whereas both *iapp* derived A-boxes contain a deviating 5´-TTAAT-3´ pentanucleotide and display affinities that are approximately five-fold weaker than Pdx1 binding to DNA_{CON}.

As a control defining the strength of non-specific double-stranded DNA interactions, we tested Pdx1 binding to the Insulin-E1 element, which lacks the TAAT sequence generally required for sequence-specific binding by homeodomains. As expected, the Insulin-E1 binds poorly, with an approximate 250-fold loss of affinity relative to DNA_{CON} under our experimental conditions. Thus, we observe that Pdx1 has a modest preference for *insulin*-derived promoter elements over those derived from the *iapp* gene, but that both are bound selectively over non-specific background DNA.

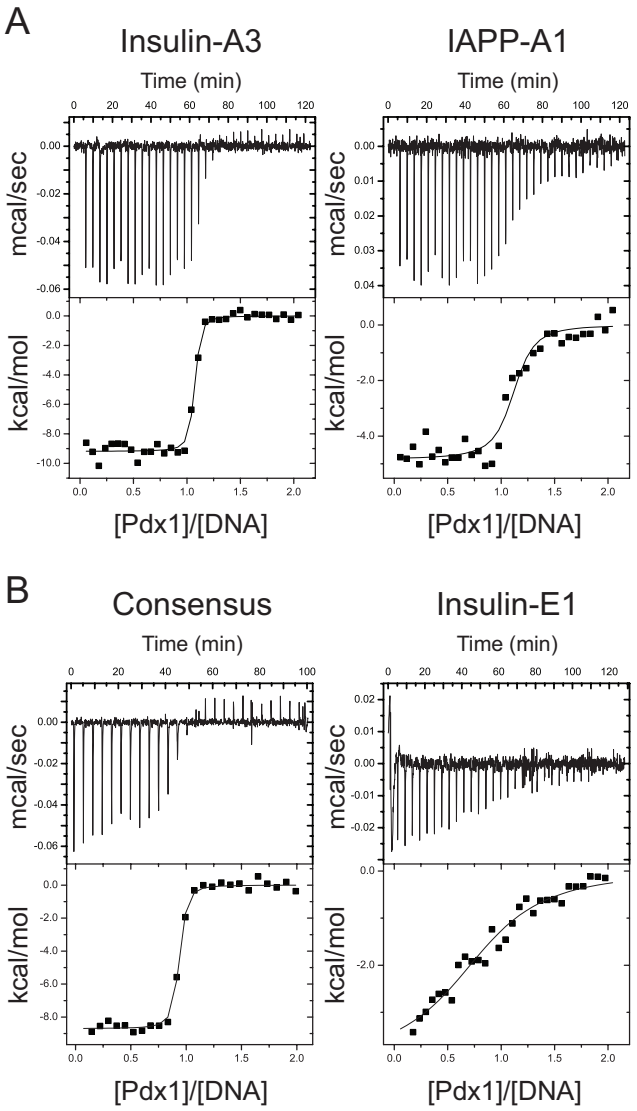

Figure **2-2**. Isothermal titration calorimetry (ITC) monitors the thermodynamics of Pdx1 interactions with natural and consensus DNA promoter elements. Representative power-response curves (top) and heats of reaction normalized to the moles of Pdx1 injected (bottom) are provided for the titration of Pdx1 into (A) Insulin-A3 DNA, (B) IAPP-A1 DNA, (C) Consensus DNA, and (D) Insulin-E1 DNA. All titrations were conducted at 298K in 100 mM cacodylate, pH 7.3, and 100 mM potassium chloride.

2.2.3 Mutational analysis using modified DNA_{CON} ligands

As shown in Table 2-1, the A-box elements of the *insulin* and *iapp* promoters differ mainly at the 5´-position relative to the core TAAT recognition element and the affinity of Pdx1 for each of these two sub-sets of A-box sequences is systematically different. Therefore, we sought to quantify Pdx1's ability to discriminate nucleotide sequence at the 5´-trinucleotide preceding the core by conducting calorimetric titrations in which we systematically varied each nucleotide's identity. For these studies, we used DNA oligonucleotides derived from the consensus sequence as 5´-CCA$\boldsymbol{N}_3\boldsymbol{N}_2\boldsymbol{N}_1$TAATGAGTTC-3´, where $\boldsymbol{N}$ in position 1, 2, or 3, respectively, is either C, G, A, or T. The thermodynamic data for the various consensus derived sequences is summarized in Table 2-2.

The results for the first position suggest that Pdx1 has a slight preference for Cyt at the 5´-position. Although the observed effect was modest (only a 2-fold change in affinity), this substitution accounts for nearly half of the 5-fold average difference observed between equilibrium binding constants for *insulin* and *iapp* derived A-box sequences (Table 2-1). Additionally, these promoters contain multiple A-boxes, often adjacent to E-boxes, as shown in Figure 2-1. Thus, we emphasize that heterotropic linkage effects involving interactions with additional binding partners recruited to the promoter by Pdx1 (and also homotropic effects involving multiple copies of Pdx1 binding to the multiple A-boxes) could readily amplify the impact of Cyt-to-Thy substitution in this position of the *iapp* promoter, relative to the *insulin* promoter sequence.

Table **2-2.** Thermodynamic parameters describing binding of Pdx1-HD to dsDNA derived from the consensus binding sequence; measured at 298K in 100 mM cacodylate, pH 7.3, 100 mM KCl. The core TAAT recognition element is shown in bold and mutations are highlighted in *italics*. All error bars on fitted ITC parameters represent standard error of the mean estimated from triplicate measurements, with fitting performed in Origin 7.0 (MicroCal, Inc.).

DNA	Forward Sequence	n	K_d (nM)	ΔH (kcal mol^{-1})
Consensus	5′-CCACTC**TAAT**GAGTTC-3′	0.9 ± 0.1	5 ± 1	-8.58 ± 0.04
Con. T1→A	5′- CCACTC***A*****AAT**GAGTTC -3′	1.0 ± 0.1	20 ± 4	-4.98 ± 0.03
Con. A2→T	5′- CCACTC**T*****T*****AT**GAGTTC -3′	1.0 ± 0.1	360 ± 50	-3.88 ± 0.04
Con. A3→T	5′- CCACTC**TA*****T*****T**GAGTTC -3′	1.0 ± 0.1	110 ± 20	-2.74 ± 0.03
Con. T4→A	5′- CCACTC**TAA*****A***GAGTTC -3′	0.9 ± 0.1	64 ± 4	-6.27 ± 0.02
1 Con. 5′C→G	5′- CCACT***G*****TAAT**GAGTTC -3′	1.0 ± 0.1	10 ± 2	-7.45 ± 0.03
1 Con. 5′C→A	5′- CCACT***A*****TAAT**GAGTTC -3′	1.0 ± 0.1	9 ± 1	-7.07 ± 0.02
1 Con. 5′C→T	5′- CCACT***T*****TAAT**GAGTTC -3′	0.9 ± 0.1	8 ± 1	-7.00 ± 0.02
2 Con. 5′T→G	5′- CCAC***G*****CTAAT**GAGTTC -3′	1.0 ± 0.0	13 ± 1	-12.6 ± 0.1
2 Con. 5′T→C	5′- CCAC***C*****CTAAT**GAGTTC -3′	0.9 ± 0.0	21 ± 1	-12.7 ± 0.1
2 Con. 5′T→A	5′- CCAC***A*****CTAAT**GAGTTC -3′	1.0 ± 0.0	22 ± 1	-12.1 ± 0.1
3 Con. 5′C→G	5′- CCA***G***TC**TAAT**GAGTTC -3′	1.1 ± 0.0	10 ± 1	-8.72 ± 0.03
3 Con. 5′C→T	5′- CCA***T***TC**TAAT**GAGTTC -3′	1.2 ± 0.0	16 ± 2	-7.76 ± 0.04
3 Con 5′C→A	5′- CCA***A***TC**TAAT**GAGTTC -3′	0.9 ± 0.0	49 ± 3	-7.87 ± 0.04
Con. 3′G→A	5′- CCACTC**TAAT*****A***AGTTC -3′	1.1 ± 0.1	11 ± 3	-4.29 ± 0.02
Con. 3′G→T	5′- CCACTC**TAAT*****T***AGTTC -3′	1.1 ± 0.1	4 ± 1	-8.86 ± 0.03
Con. 3′G→C	5′- CCACTC**TAAT*****C***AGTTC -3′	1.1 ± 0.1	30 ± 3	-9.08 ± 0.02

The *insulin* and *iapp* sequences vary substantially in the 5´-region relative to the TAAT core. Based on the binding results of Pdx1 with position 1 consensus variants, we hypothesized that Pdx1 is able to discriminate sequence upstream the core binding site. Thus, we varied the nucleotide identity in position 2 to determine the effects of DNA binding by Pdx1. A slightly greater effect is observed in the binding results when the nucleotide in position 2 was systematically mutated. Altering the nucleotide identity to a Gua, Cyt, or Ade in this position decreased Pdx1's binding affinity to 13 nM, 21 nM, and 22 nM, respectively. Surprisingly, the binding affinities of each of these sequences were weaker than that of the consensus sequence, suggesting Pdx1 also interacts with the nucleotide in position 2. Thus, Pdx1 is preferential towards a Thy in position 2 based on the higher affinity of 5 nM.

Based on the co-crystal structure of Pdx1 with DNA, only a couple residues interact with bases in the minor groove, especially with the first two bases in the core binding site. However, calorimetric results for binding of consensus-derived mutants (position 1 and position 2 variants) yield results that show Pdx1 discrimination at these two positions. When comparing the *insulin* and *iapp* promoter derived sequences, the *insulin* sequences contain a Cyt at position 3, while the *iapp* promoter derived sequences both contain an Ade in this position. Does nucleotide identity at this position influence binding by Pdx1?

The calorimetric results for the third position are unanticipated, showing a marked decrease in binding affinity from 2-fold, 3-fold, and 10-fold weaker based on nucleotide identity of Gua, Thy, and Ade, respectively. As opposed to a Cyt in the wild-type consensus sequence, a Gua in this position only modestly perturbs the interaction. Interestingly, a Cyt or Gua in position 3 is more favorable for binding and slightly more favorable energetically than an Ade or Thy base in this position. Binding by Pdx1 to Thy results in a 3-fold reduction in binding affinity, while a significant, that is 10-fold, decrease in binding affinity to Ade in position 3 is observed for Pdx1. While some of the binding affinity variation between *insulin* and *iapp* sequences comes from the Cyt to Thy

substitution at position 1, most of the variation in binding may be attributable to the Ade in position 3 in the *iapp* sequences. The results certainly show that Pdx1 discriminates nucleotide sequence in the 5´-region preceding the core site. This is one of the first studies showing the ability of a homeodomain protein to distinguish nucleotide identity upstream the core binding site. Illuminating the mechanism by which Pdx1 differentiates nucleotide sequence remains an area of interest.

In addition to determining the binding affinity in a label-free assay, ITC provides direct determination of binding enthalpies, which can provide insight into the molecular mechanisms of interactions. Pdx1 binding to Cyt in the 5´-position is enthalpically more favorable than Gua, Ade, or Thy in this position (Table 2-2), which is consistent with the systematic difference in binding enthalpy of *insulin* versus *iapp* derived sequences (Table 2-1) and with the differential electrostatic stabilization of Arg-150 in the minor groove observed in molecular dynamics simulations of Pdx1 in complex with each of these four sequences (*vide infra*).

The significance of the Cyt-to-Thy substitution discussed above has the potential to impact relative recruitment of Pdx1 to the *insulin* and *iapp* promoters and underscores the need to explore Pdx1 sequence preferences rigorously. Further motivating our research, a systematic study of the engrailed homeodomain by Ades *et al*. demonstrated the effects of modifying bases in the core and flanking regions of its preferred binding sequence, revealing regions where mutation decreased the binding affinity and modified functional outcomes.[29] Therefore, we continued our study with Pdx1 titrations against variants of the DNA_{CON} sequence intended to disrupt the TAAT motif (summarized in Table 2-2).

In the co-crystal structure of Pdx1 bound to DNA, the first and fourth bases in the TAAT core sequence are only contacted by a single residue. Moreover, the only direct contact to the fourth base pair is a van der Waals interaction in the minor groove mediated by Ile-192, which led Longo *et al*. to speculate that Pdx1 may be tolerant of DNA mutations in this position.[22] Through ITC, we observe that when the Thy bases at position 1 and 4 are inverted to Ade, the Pdx1 binding affinity weakens by

4-fold and 12-fold, respectively. Interestingly, inverting the base pair at position 1 of the TAAT element results in a binding enthalpy close to that of the *iapp* promoter-derived sequences and much smaller in magnitude than those observed for the *insulin* promoter-derived and wild-type consensus sequence. Moving into the interior of the TAAT core site, the crystal structure reveals that significantly more residues in helix-3 and in the disordered N-terminal arm contact the bases in positions 2 and 3. Unsurprisingly, a more pronounced effect on the K_d is observed when positions 2 and 3 are inverted from Ade to Thy, weakening the affinity by approximately 75-fold and 20-fold, respectively.

To complete our mutational analysis, we systematically varied the nucleotide identity at the 3′-position relative to the TAAT element, because prior reports suggest that Pdx1 exhibits sequence bias for Gua or Thy at this location.[18; 19] The Pdx1-DNA co-crystal structure demonstrates that Met-199 makes hydrophobic contacts with the Cyt base of the 3′ G·C base pair on the strand opposing the TAAT element. Similar interactions would be possible with an Ade base in a T·A base pair, suggesting a mechanism for the Gua/Thy preference predicted at the 3′-position. Our ITC results corroborate these findings, showing a 2- to 6-fold enhancement in affinity for Gua or Thy in the 3′-position, as compared to the binding affinity observed for Ade or Cyt in this position (Table 2-2).

In summary, we find that the TAAT element is indispensible for sequence-specific Pdx1 binding, as predicted by the many interactions between Pdx1 helix-3 and the DNA nucleobases in the co-crystal structure. Additionally, nucleobase-specific interactions in the 5′- and 3′-flanking sequences predicted from establishment of the 5′-CTAAT(T/G) -3′ consensus are confirmed by our studies, although we find that the quantitative contributions from the 5′ C·G and 3′ T·A or G·C base pair are small in magnitude. Significantly, discrimination between a 5′ C·G base pair, seen in human *insulin* promoter sequences, and a 5′ T·A base pair, seen in human *iapp* promoter sequences, is established. As with many homeodomains that recognize nucleotide identity in the region proximal to the 5′-side of the TAAT motif, recognition of the 5′-pair by Pdx1 appears from the crystal structure

to be mediated through the minor groove of the duplex DNA, where there is a paucity of chemical information capable of defining nucleotide sequence, as compared to what would be made available through major groove interactions. Therefore, the remainder of the work presented here will aim to provide a mechanistic explanation for the thermodynamic preferences we have quantified by ITC with respect to nucleotide composition on the 5´-side of the TAAT motif.

2.2.4 The Role of Arg150 in DNA-Binding and Sequence Selectivity

Arg-150 is strictly conserved in all homeodomain sequences. In the canonical structure of homeodomain proteins, the disordered N-terminal arm inserts into the minor groove of DNA enabling Arg-150 to interact directly with bases. Due to the local negative charge and narrow width of the minor groove, arginine and lysine residues are the principal candidates for making contacts within DNA minor grooves. These basic amino acids hydrogen bond with bases in the minor groove, primarily to N3 in purines and O2 in pyrimidines.[30] In the case of homeodomain proteins, Arg-150 interacts with the first base in the core binding site by hydrogen bonding. Because arginine plays a significant role in the DNA binding of homeodomains, a mutational analysis was performed to determine the effects of removing the positive charge and also substituting the positive charge with a similarly charged amino acid.

Removing the positive charge at position 150 was detrimental to the overall binding of the protein. Several conserved residues in the homeodomain sequence are responsible for sequence selectivity, with most residues in the DNA recognition helix. Based on the co-crystal structure of Pdx1, Lys-147 and Arg-150 directly interact with DNA via hydrogen bonding. Calorimetric data shows that Pdx1-HD R150A was unable to selectively bind to any DNA sequence, rather the interaction was in the non-specific binding regime (Table 2-3). Pdx1-HD interacted with Ins E1 non-specifically with an affinity of 1300 nM; similarly the mutant Pdx1-HD R150A interacted with the

DNA panel non-specifically. Recorded affinities of the R150A mutant for near-consensus A-box sequences averaged 3200 nM. Interestingly, there was no detectable binding of Pdx1-HD R150A with Ins E1, which does not contain the TAAT core binding site. Although, the R150A mutant still contains the residues that make sequence specific contacts within the major groove of DNA, the mutant was unable to bind at all, underscoring the importance of Arg-150 in DNA-binding. These findings are consistent with a previous report on the Engrailed homeodomain, where binding was reduced approximately 400-fold when Arg-150 was mutated to an alanine.[29]

Not only is Arg-150 important for interacting with the first two bases in the core binding site, but it is vital for the binding of the protein. As shown in the electrostatic potential from MD simulations (section 2.2.7), the insertion of Arg-into the minor groove helps to neutralize a strong negatively charged pocket. Arginine side chains carry a positively charged guanidino group that is also capable of bi-directional hydrogen bonding interactions, thus enabling arginine to bridge two bases. On the other hand, the lysine side chain possesses a positively charged amino group that can substitute for arginine's charge-screening effects, but lacks arginine's hydrogen bonding geometry. For this reason, it is interesting to investigate whether R150K substitution is able to retain wild type binding affinity and sequence selectivity. The binding affinity of Pdx1-HD R150K was approximately 30-fold weaker and sequence selectivity was dramatically affected. The binding enthalpies of Pdx1-HD R150K were approximately 2-fold lower than those of Pdx1-HD, strongly suggesting the R150K mutant does not contribute similarly to the overall enthalpy of binding. This finding is striking, in that one would expect lysine to behave similarly electrostatically as arginine due to the positively charged side chain. However, sequence selectivity, binding affinity, and enthalpy are compromised when lysine occupies position 150. Collectively, this data demonstrates the importance of the evolutionarily conserved Arg-150 (position 5 in the canonical homeodomain sequence) in binding affinity and sequence specificity.

Table **2-3** Thermodynamic parameters describing binding of Pdx1-HD R150A and R150K mutants to dsDNA derived from the consensus binding sequence; measured at 298K in 100 mM cacodylate, pH 7.3, 100 mM KCl. The core TAAT recognition element is shown in bold and mutations are highlighted in red. All error bars on fitted ITC parameters represent standard error of the mean estimated from triplicate measurements, with fitting performed in Origin 7.0 (MicroCal, Inc.).

		Homeodomain		**Homeodomain R150A**		**Homeodomain R150K**	
Sample Name	**Sequence**	**K_d (nM)**	**ΔH (kcal/mol)**	**K_d (nM)**	**ΔH (kcal/mol)**	**K_d (nM)**	**ΔH (kcal/mol)**
Insulin A3	5′-AGACTC**TAAT**GACCCG-3′	4 ± 1	-9.6 ± 0.1	1670 ± 70	-4.46 ± 0.06	90 ± 10	-5.19 ± 0.03
Consensus	5′-CCACTC**TAAT**GAGTTC-3′	5 ± 1	-8.6 ± 0.1	3440 ± 150	-4.24 ± 0.09	140 ±10	-4.61 ± 0.02
Consensus 5′G	5′- CCACT*G***TAAT**GAGTTC -3′	10 ± 2	-7.45 ± 0.05	4630 ± 180	-5.9 ± 0.1	120 ± 10	-5.32 ± 0.03
Consensus 5′A	5′- CCACT*A***TAAT**GAGTTC -3′	9 ± 1	-7.07 ± 0.04	3350 ± 230	-3.6 ± 0.1	210 ± 20	-3.86 ± 0.03
Consensus 5′T	5′- CCACT*T***TAAT**GAGTTC -3′	8 ± 1	-7.00 ± 0.04	2930 ± 170	-2.79 ± 0.06	70 ± 10	-4.09 ± 0.03
Insulin E1	5′-AGCCATCTGCCGACCC-3′	1300 ± 300	-2.7 ± 0.2	NDB[a]	NQD[b]	6900 ± 200	NQD

[a] NDB: There was no detectable binding for this DNA sequence. [b] NQD: The binding enthalpies were not quantifiable due to the weak binding or undetectable binding.

2.2.5 Long timescale molecular dynamics simulations

Placing the thermodynamic data presented above in the context of the co-crystal structure of Pdx1 bound to a consensus derived DNA duplex suggests that penetration of the Arg-150 side chain into the minor groove is responsible for imparting 5´ nucleotide specificity (Figure 2-1). While the crystallographic orientation of Arg-150 explains its contribution to recognizing the first position of the core TAAT element, it provides no clear evidence for how Arg-150 may mediate sequence discrimination at the 5´-position relative to TAAT, because no nucleobase-specific contacts are made to this base pair.[22] In all likelihood, the crystallographic conformation of Arg-150 reflects one of many thermally accessible conformations. Therefore, experiments designed to enumerate more fully the conformational states accessible to this key residue – as well as those of the remainder of the N-terminal arm and homeodomain – could provide insight into the mechanism of sequence discrimination imparted by the N-terminal arm.

Most experimental methods, such as the NMR spin relaxation studies also reported here, provide a temporally averaged description of the dynamics in biomolecular systems. Therefore, computational approaches that enable time-resolved sampling of conformational ensembles reflect an important complement to experimental studies. Recent advances in both computing power and force field quality have made computer simulation of dynamic regions like the Pdx1 N-terminal arm on meaningful timescales feasible.[31; 32] Therefore, we chose to calculate molecular dynamics trajectories of Pdx1, starting from the co-crystal structure, based on the hypothesis that thermally accessible fluctuations away from the crystallized conformation of the tail would elucidate the atomic-scale interactions responsible for our observed binding thermodynamics. In order to assure sufficient sampling to test our hypothesis, we have conducted our simulations using the Anton machine, which was purpose-built to enable long-timescale molecular mechanics

calculations for biomolecular systems.[33] We have calculated 5 μs simulations for Pdx1-DNA complexes including each of the four 5´-position variants used in our ITC studies, as well as a 5 μs simulation of apo-Pdx1, resulting in 25 μs of total sampling. The results, which we validate through comparison with experimental NMR data, provide a reasonable atomic-scale mechanism for nucleobase-sequence identification at the 5´-position, relative to the TAAT core, by Arg-150.

The stability of the homeodomain and homeodomain-DNA complex during our 5 μs Anton simulations is shown by the root mean squared deviation from the starting conformation, displayed as a function of time in Figure 2-3. Both apo-Pdx1 (grey) and the Pdx1-DNA complex (black) were found to be remarkably stable overall on this extended timescale, with the C-terminal end of helix-3 being a major exception to that trend. There is precedent to support this result, as the NMR solution structure of the homeodomain Antp-DNA complex (pdb 1ahd) displays significant disorder in the same region of helix-3.[34] Still, we were concerned that the unraveling of helix-3 in our simulations may have been an artifact, and so we sought to quantify the extent of helicity in the same region of apo-Pdx1 and the Pdx1-DNA_{CON} complex experimentally.

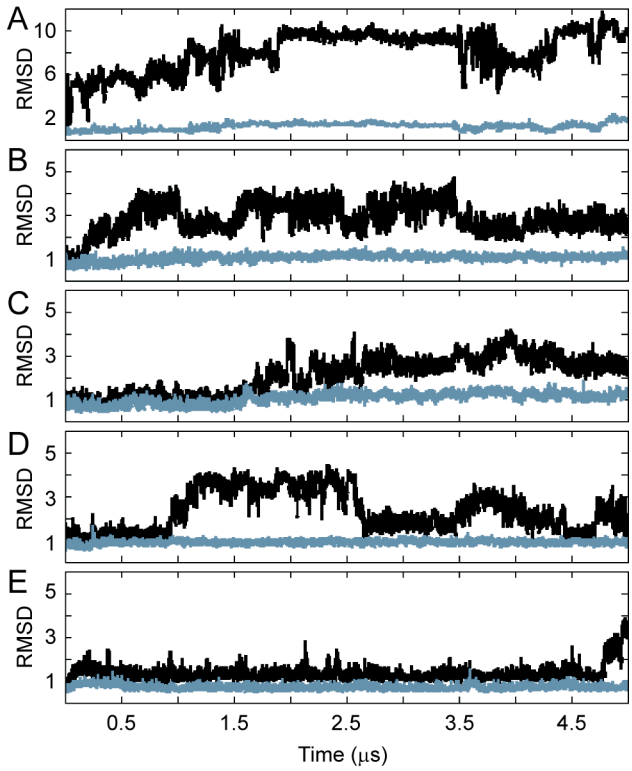

Figure **2-3.** Pdx1 root-mean-squared deviation (RMSD) from the crystal structure during the five-microsecond production molecular dynamics calculations performed on Anton. Results are shown for (A) apo-Pdx1, (B) Pdx1-DNA$_{CON}$(CTAATG), (C) Pdx1-DNA(ATAATG), (D) Pdx1-DNA(GTAATG), and (E) Pdx1-DNA(TTAATG). For each trajectory, the Pdx1 RMSD (Å) is shown for all backbone heavy atoms (black) and for the backbone heavy atoms of residues 155-199 only (grey). Data points are reported once per 200 ps of simulation time.

Backbone NMR chemical shifts are a robust indicator of polypeptide secondary structure and are therefore the ideal metric to assess similarity in the average helicity of helix-3 in our simulations and *in vitro*. As summarized in Figure 2-1B, we have acquired NMR chemical shifts for the backbone of both apo-Pdx1 and the Pdx1-DNA$_{CON}$ complex (deposited in the BMRB with ID numbers 19227 and 19228, respectively). Secondary structure propensity (SSP) is widely used method for utilizing chemical shifts to predict the structure of proteins; α-helix and β-strand structures, as opposed to random coil structures, are predicted based on positive or negative SSP values, respectively. Notably, values greater than or equal to 2 in absolute magnitude signal stable

secondary structure.[35] Importantly, the N-terminal arm of Pdx1 is shown by SSP analysis to be disordered in apo-Pdx1 for those residues having assignments (Figure 2-4A), as expected. Apo-Pdx1 also displays three clear regions of α-helical structure, but the boundaries of these structures are only consistent with the co-crystal structure for helix-1 and helix-2; the C-terminus of helix-3 shows significant fraying in the SSP data, which is qualitatively consistent with the predictions from our MD simulations.

Similarly to the data for apo-Pdx1, the experimental chemical shifts from the Pdx1-DNA_{CON} complex yield SSP values that indicate a consistent overall structure for the homeodomain *in vitro* and in our MD simulations. First, while the N-terminal arm is ordered in the complex, assignment of regular α-helical or β-strand structure is not merited by SSP (Figure 2-4B). Second, assignment of all three α-helices is similar between the experimental and MD chemical shifts, although fraying of helix-3 is more pronounced in the MD simulation than in our *in vitro* experiments. This result is consistent with what we also observed for apo-Pdx1. Taken together, the chemical shift analysis suggests that the average conformations of apo-Pdx1 and the Pdx1-DNA_{CON} complex in our MD simulations are representative of the solution averages. The existence of fraying in the C-terminus of helix-3 and the known dynamic state of the N-terminal arm suggest that additional analysis of the fluctuations about these average structures will also yield insight into the structural ensembles and DNA-binding mechanism of Pdx1.

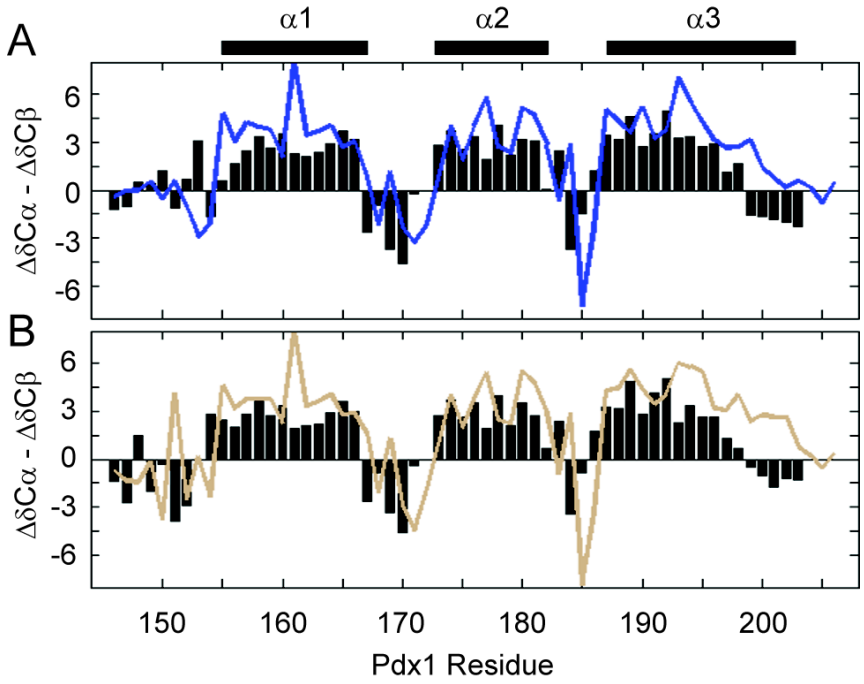

Figure **2-4**. Secondary $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ chemical shifts indicate the boundaries of secondary structure in (A) apo-Pdx1 and (B) the Pdx1-DNA_{CON} complex. Plotting the difference between $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ secondary chemical shifts indicates the presence of an α-helix when long stretches of positive values are encountered. The secondary structure from the co-crystal of Pdx1 with a consensus DNA duplex is represented by bars above the figure for comparison. In both panels, the experimental secondary shift difference is reported by a colored line, whereas the predicted values resulting from the Anton MD trajectories is represented by black bars.

2.2.6 Backbone dynamics of Pdx1 in the apo- and DNA-bound states

If the overall folding features we observed in our Anton trajectories are reasonable, then the enhanced disorder in the N-terminal arm and the C-terminus of helix-3 should manifest in larger amplitude fluctuations of the backbone conformation in these regions, relative to the temporally stable portions of the protein found in the folded core of the homeodomain. In particular, enhanced dynamics on the ps-ns timescale should be observed for the disordered N-terminal tail and the frayed end of helix-3. Such dynamics are measurable through NMR spin relaxation methods.[36; 37] Importantly, these dynamics can also be accessed through analysis of the fluctuations recorded in our MD simulations,[38] allowing additional cross-validation of the

experimental and computational data sets. Therefore, we have collected ^{15}N-NMR spin relaxation measurements on a 500 MHz spectrometer for apo-Pdx1 and the Pdx1-DNA$_{CON}$ complex, which are summarized in Figure 2-5. The raw data indicate a picture that is consistent with the chemical shift analysis reported above, where the N-terminal arm of Pdx1 appears to be relatively static in the complex and the C-terminus of helix-3 appears to be more dynamic in the apo-state.

In order to better quantify the dynamics we observed, we performed modelfree analysis[39] of the ^{15}N-NMR spin relaxation data for both apo-Pdx1 and the Pdx1-DNA$_{CON}$ complex. Data analysis began with determination of the global tumbling properties of both apo- and DNA-bound Pdx1. Application of an axially symmetric diffusion tensor was merited by the data, resulting in an effective isotropic correlation time (τ_c) of 5.8 ns and 12.6 ns for apo-Pdx1 and the Pdx1-DNA$_{CON}$ complex, respectively; the axial ratio of the diffusion tensors was 1.28 and 1.29, respectively. NMR spin relaxation for several other homeodomain proteins has been documented in the literature, including for Pitx2, MATa1, and vnd/NK-2.[40; 41; 42] In two of these three unbound state cases, the reported τ_c was greater than the 5.8 ns correlation time we observed for apo-Pdx1. MATa1 homeodomain, which had a very similar mass to our Pdx1 construct, yielded a similar correlation time of 5.09 ns. The elevated correlation time observed for the Pdx1-DNA$_{CON}$ complex is consistent with the complex's increased mass and the magnitude we observed is comparable to the 11.5 ns correlation time reported for DNA-bound vnd/NK-2.[41] Overall, our spin relaxation data, as well as dynamic light scattering results (not shown), are consistent with a monodisperse and monomeric state of both apo-Pdx1 and the Pdx1-DNA$_{CON}$ complex in solution.

In order to better quantify the dynamics we observed, we performed modelfree analysis[39] of the ^{15}N-NMR spin relaxation data for both apo-Pdx1 and the Pdx1-DNA$_{CON}$ complex. Data analysis began with determination of the global tumbling properties of both apo- and DNA-bound Pdx1. Application of an axially symmetric diffusion tensor was merited by the data, resulting in an effective isotropic correlation time (τ_c) of 5.8 ns and 12.6 ns for apo-Pdx1 and the Pdx1-

DNA_{CON} complex, respectively; the axial ratio of the diffusion tensors was 1.28 and 1.29, respectively. NMR spin relaxation for several other homeodomain proteins has been documented in the literature, including for Pitx2, MATa1, and vnd/NK-2.[40, 41, 42] In two of these three unbound state cases, the reported τ_c was greater than the 5.8 ns correlation time we observed for apo-Pdx1. MATa1 homeodomain, which had a very similar mass to our Pdx1 construct, yielded a similar correlation time of 5.09 ns. The elevated correlation time observed for the Pdx1-DNA_{CON} complex is consistent with the complex's increased mass and the magnitude we observed is comparable to the 11.5 ns correlation time reported for DNA-bound vnd/NK-2.[41] Overall, our spin relaxation data, as well as dynamic light scattering results (not shown), are consistent with a monodisperse and monomeric state of both apo-Pdx1 and the Pdx1-DNA_{CON} complex in solution.

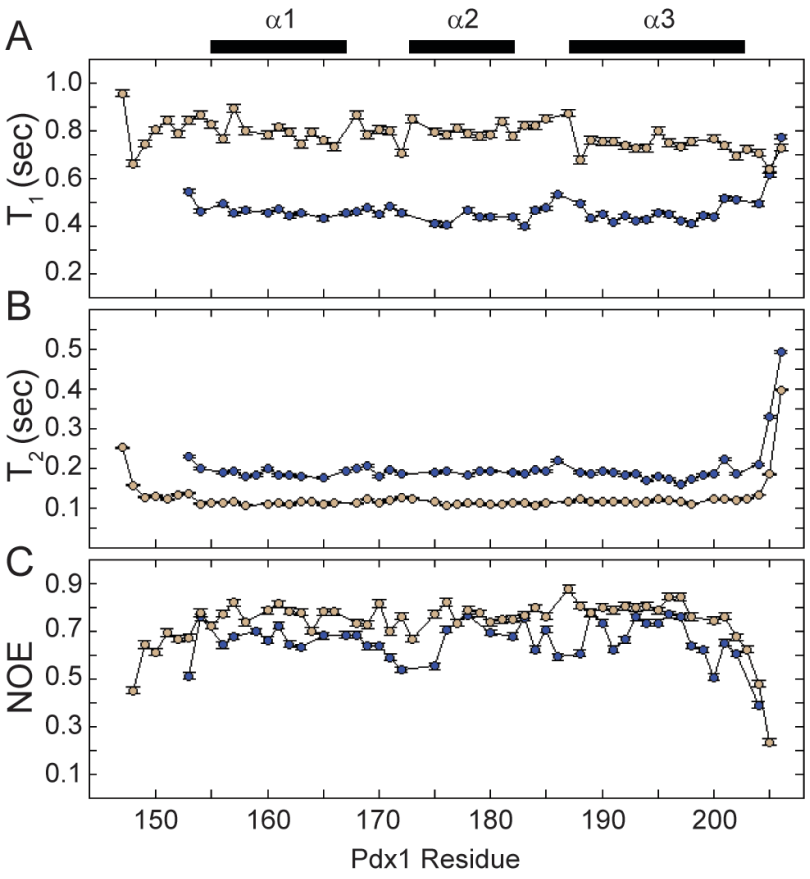

Figure **2-5.** Backbone ^{15}N-T_1, T_2, and ^{1}H,^{15}N-NOE measured on a 500 MHz NMR spectrometer. (A) ^{15}N-T_1, (B) ^{15}N-T_2, and (C) ^{1}H,^{15}N-NOE are reported for apo-Pdx1 (blue) and the Pdx1-DNA_{CON} complex (tan). Experimental uncertainties in the measured parameters are indicated as error bars, which often do not exceed the size of the markers on the plot. The secondary structure from the co-crystal of Pdx1 with a consensus DNA duplex is represented by bars above the figure.

Modelfree analysis of the experimental spin relaxation data and iRED analysis[43] of the molecular dynamics simulations yields qualitatively similar order parameter profiles (Figure 2-6). Overall, the order parameters suggest increased rigidity of Pdx1 in complex with DNA, although two specific regions merit additional discussion. Consistent with the known biochemistry of homeodomains, the Pdx1-DNA co-crystal structure suggests that the N-terminal residues of the

Pdx1 homeodomain provide additional DNA sequence specificity by contacting the minor groove of DNA. In the apo-state, several of the N-terminal residues are not observed in the $^{1}H,^{15}N$-HSQC due to the fast exchange of their amide protons with solvent (confirmed in experiments conducted at pH 6.5, in which these resonances are restored; data not shown). In the DNA-bound state, the N-terminal residues become protected from fast solvent exchange through their interactions with the minor groove of the DNA and therefore are observed in the 2D-spectra. Order parameters for the disordered N-terminal tail of the homeodomain are generally quite high in the Pdx1-DNA_{CON} complex, suggesting that the structure of this region, though irregular, is temporally stable in the complex. Additionally, the MD derived order parameters for the N-terminal arm of Pdx1 in complex with DNA_{CON} are in nearly quantitative agreement with the experimental results, which further supports the conclusion that the N-terminal arm interacts stably with the minor groove of the DNA in solution. Contrasting this data with the MD simulated order parameters of the apo-state suggests the N-terminal arm of apo-Pdx1 (which cannot be studied experimentally due to solvent exchange) is highly dynamic on the ps-ns timescale.

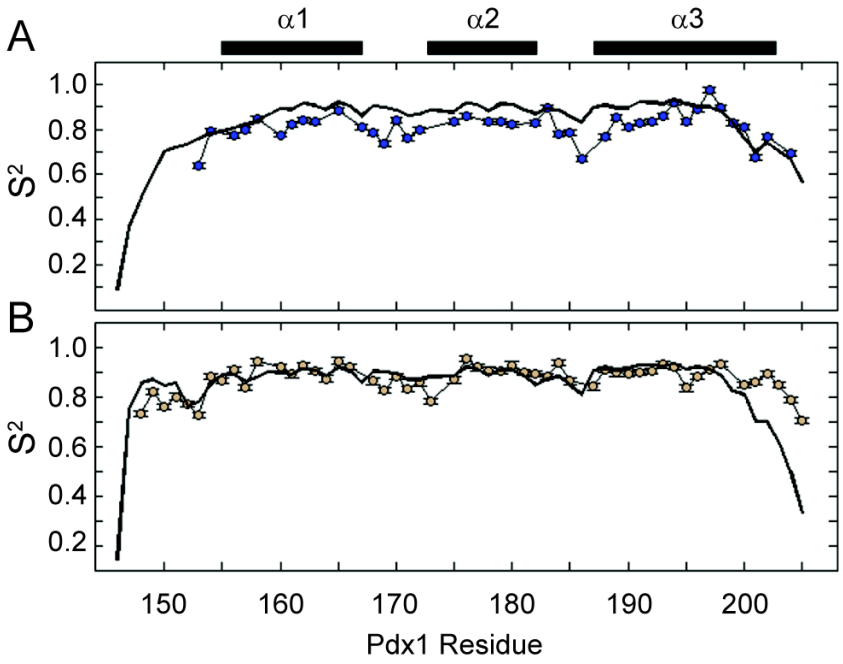

Figure **2-6**. Backbone amide generalized order parameters (S^2) as a function of residue number for (A) apo-Pdx1 and (B) the Pdx1-DNA_{CON} complex. In both panels, the experimental order parameters are represented by colored circles attached with thin black lines, while those calculated from the Anton MD trajectories are represented as a thick black line. Experimental uncertainties in the order parameters are indicated as error bars, which often do not exceed the size of the markers on the plot. The secondary structure from the co-crystal of Pdx1 with a consensus DNA duplex is represented by bars above the figure.

In addition to the increased order of the N-terminal tail in the bound-state, the spin relaxation data reveal qualitative differences in the dynamics of the C-terminal turn of helix-3 in the apo- and DNA-bound states, as compared to the dynamics of the core secondary structure elements. Extensive fraying of the C-terminal turn of helix-3 during both the apo-Pdx1 and Pdx1-DNA_{CON} simulations manifests as lower order parameters in this region (Figure 2-6). As with the chemical shift data, the magnitude of unwinding predicted by simulation is in good agreement with the experimental values for apo-Pdx1 (Figure 2-6A). In addition, the enhanced fast timescale dynamics of the C-terminal turn of helix-3 suggested by the order parameter profile are corroborated by the raw heteronuclear NOE data seen in Figure 2-5C. Although the magnitude of

fraying for the Pdx1-DNA_{CON} complex reported by decreased order parameters in the simulations is not quantitatively consistent with experiment (Figure 2-6B), the onset of fraying occurs at a similar location in both data sets. In these data, the dynamics of helix-3 begin to increase near Arg-198 in the unbound state and with Met-199 in the bound state. This correlates well with the simulated and experimental SSP calculations for both states and with the known role of Met-199 in forming interactions with the nucleobase at the 3´-position relative to the core TAAT sequence. Taken together, our MD and NMR data suggest that the C-terminal end of helix-3 is less stable than indicated in the crystal structure, although the portion of the helix that forms the DNA-interaction surface is well ordered in both the apo- and DNA-bound state. In summary, the overall structural and dynamic description of apo-Pdx1 and the Pdx1-DNA_{CON} complex in our MD simulations compares favorably with experimental NMR data, suggesting that theses trajectories are well suited to generate a mechanistic hypothesis for the origin of sequence specificity observed in our ITC study.

2.2.7 Implications for insulin and iapp promoter interactions

Sequences similar to the consensus 5´-CTCTAAT(T/G)AG-3´ recognition element are commonly found in multiple copies throughout the proximal promoter regions of β-cell associated genes. Because many homeodomains bind a TAAT core recognition site, interaction between peripheral nucleotide sequences and upstream amino acids may play a key role in providing sequence discrimination. Importantly, many early reports of promoter binding by Pdx1 in biochemical assays relied heavily on rat/mouse promoter sequences and did not address the effects of critical variations between these and human promoters.[10; 28; 44; 45] The rat *insulin* promoter is significantly dissimilar to the human promoter, with changes to the A3-box being particularly pronounced.[19] Of relevance to the present study, there is a substitution of the 5´

position relative to the TAAT core in the rat and mouse insulin A1 element (5′-TTAAT-3′), compared to the more canonical 5′-CTAAT-3′ sequence seen in human. In other words, near the core TAAT element, some A-boxes in the rat *insulin* promoter appear more similar to the A-boxes of the human *iapp* promoter than they do to human *insulin* promoter elements. Given that we have shown these substitutions are thermodynamically significant, it is also important to establish a plausible model for the molecular origin of sequence recognition at this 5′-position.

In the crystal structure of Pdx1 bound to a consensus DNA sequence, only the side chain of Arg-150 penetrates into the minor groove of the DNA duplex and makes direct contact with the nucleobases in the 5′-proximal region. By ITC, we have shown that Pdx1 has a modest preference for a C·G base pair in the 5′-position relative to the core motif. In order to provide a molecular rationale for this observation, we have extended our set of microsecond timescale MD simulations to include three trajectories where the 5′ base pair is systematically altered to T·A, G·C, and A·T. Simulation of the complex formed between Pdx1 and the consensus DNA containing a 5′ C·G base pair shows that Arg-150 is capable of achieving orientations that permit simultaneous hydrogen bonding to the O2 atom of the Thy in position 1 of the TAAT core motif and N3 of the guanine nucleobase on the opposing strand of the C·G base pair in the 5′-position relative to the core motif. While this bridging interaction is remarkably stable in the Pdx1-DNA(CTAAT) simulation, no such stable interaction involving the nucleobases of the 5′-pair is observed for any of the other three Watson-Crick pairing combinations. For visual aid, the snapshot of each MD simulation recorded after 3.5 μs of production dynamics is shown in Figure 2-7, with the zoom set such that interactions between Arg-150, the 5′-pair, and the Thy in position 1 (Thy-1) of the TAAT motif are easily seen. In each panel, Arg-150, Thy-1, and the base occupying the opposing strand in the 5′-pair are shown in color, while the remainder of the DNA is represented in white.

In the crystallographic orientation, the only direct hydrogen bonds between a DNA nucleobase and the guanidino group of Arg-150 are formed to the O2 atom of Thy-1, as illustrated in Figure 2-7A. This result contrasts with our ITC data, which show a subtle but clear preference for a C·G base pair in the 5′-position relative to the TAAT core. If Arg-150 is to sense the nucleotide identities in this pair, it is likely that thermally accessible fluctuations are able to drive reorientation of Arg-150 hydrogen bonding groups towards the 5′-pair. This is precisely what we observe; approximately 1 μs into the simulation, Arg-150 reorients such that it simultaneously hydrogen bonds to the O2 of Thy-1 (through the guanidino NH1 group) and the N3 of the 5′-Gua on the opposing strand (through the guanidino NH2 group). Once adopted, this orientation remains stable until fluctuations briefly drive the orientation back to a conformation more similar to the crystallographic orientation at around 2.5 μs of simulation time; but this transition is short lived and the bridging orientation is quickly resumed for the remainder of the simulation. The snapshot recorded at 3.5 μs of simulation time (shown in Figure 2-7B) is representative of the most stable orientation and illustrates the bridging interaction clearly.

From the 5′-CTAAT-3′ simulation alone, it is tempting to speculate that hydrogen bonding between the Arg-150 side chain and the N3 atom of any purine on the opposing strand, in the 5′-position relative to the TAAT core, is the primary determinant of the nucleotide preference of Arg-150. However, if this were true, a T·A pair in the 5′-position, as is observed in the *iapp* gene promoter, should provide equally strong binding – and yet our ITC data indicated this is not the case. Much to our surprise, in the Pdx1-DNA(TTAAT) simulation, Arg-150 retains a conformation similar to that of the co-crystal structure, never making temporally stable contact with the 5′ T·A base pair (as represented in Figure 2-7C).

Given the unexpected nature of the results from our calculations with a 5′-TTAAT-3′ sequence, we next sought to determine the outcome of flipping the 5′-pair, producing the sequences 5′-GTAAT-3′ and 5′-ATAAT-3′. In these trajectories, Arg-150 appears to

preferentially hydrogen bond to the exocyclic O2 of the pyrimidine in the strand opposing the TAAT motif, to the exclusion of interactions with Thy-1. In this case, the loss of stability is easy to rationalize from the structures in the trajectories, because Arg-150 is largely expelled from the minor groove in order to accommodate this hydrogen bond, which fluctuates between forming and breaking multiple times over the course of both trajectories. The representative structures at 3.5 μs of simulation time for the 5′ G·C and 5′ A·T mutants (shown in Figure 2-7D, E) illustrate this conformational transition clearly. Recall that in the ITC studies of binding to DNA_{CON}, bearing mutations at the 5′-position, the binding enthalpy was less favorable for each of the three mutant sequences, as compared with the enthalpy observed for binding to the wild-type 5′-CTAAT-3′ consensus. Partial expulsion of Arg-150 from the minor groove in the presence of the 5′ G·C or 5′ A·T base pairs, with only transient hydrogen bonding to the pyrimidine O2, is consistent with the loss of stabilizing binding enthalpy. On the other hand, binding to the *iapp*-type consensus element containing a 5′-TTAAT-3′ sequence, and binding to *iapp*-derived native sequences, also occurs with a significantly less favorable binding enthalpy, as compared to *insulin* promoter or DNA_{CON} sequences. This key result is more challenging to rationalize based solely on hydrogen bond geometry.

These data suggested to us that an additional factor, beyond stable penetration into the minor groove and hydrogen bonding to Thy-1, is important for determining 5′-pair specificity. Noting that in our simulations of Pdx1-DNA(CTAAT), Arg-150 reoriented to interact with the Gua-N3 in the opposing strand of the 5′-position, but not the Ade-N3 when the 5′ T·A mutation was made. Therefore, we investigated whether the nature of the minor groove itself is different between consensus DNA harboring the preferred 5′ C·G and the less favorable 5′ T·A. As both of these base pairs are sterically similar and present a similar pattern of hydrogen bond donors and acceptors in the minor groove, we turned to calculations of the surface electrostatics in the minor groove as a possible explanation for our binding data.

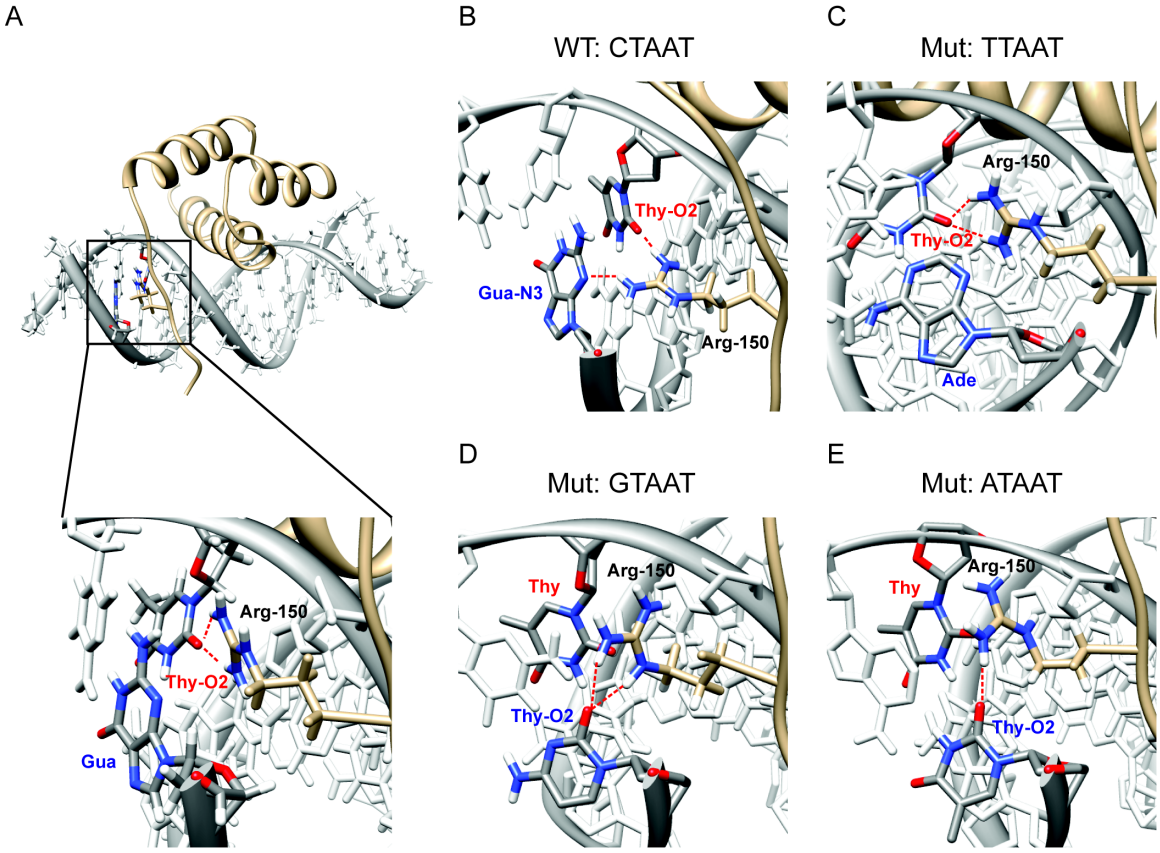

Figure **2-7**. Reorientation of the Arg-150 side chain during the Anton MD simulations rationalizes the exclusive preference for a 5′ C·G base pair in sequences observed in the ITC study. For clarity, the Thy base from the TAAT sequence and the nucleobase on the opposing strand in the 5′-flanking base pair are identified in color in all panels, whereas all other nucleobases are shown in white. (A) The co-crystal structure of Pdx1 bound to consensus DNA shows Arg-150 to be the only residue capable of making direct contact with nucleotides 5′ to the core TAAT element. Contrary to the observation of selectivity for a C·G base pair observed by ITC, the co-crystal only shows Arg-150 in direct contact with the first T of the core element – no contact to the C·G base pair is made (inset). (B) During the Anton MD trajectory, Arg-150 rapidly reorients to simultaneously hydrogen bond with the Thy-O2 of the first position in the TAAT core element and with the Gua-N3 on the opposing strand in the 5′ C·G base pair. (C) Arg-150 retains a conformation similar to that of the co-crystal structure when Pdx1 is simulated in complex with the TTAAT mutant DNA, never making temporally stable contact with the 5′ T·A base pair. In both the (D) 5′ G·C and (E) 5′ A·T mutant trajectories, interactions of the Arg-150 side chain with nucleobase hydrogen-bond acceptors in the minor groove of the DNA duplex are especially unstable. In all cases, the snapshot recorded after 3.5 μs of production dynamics displays the predominant orientation of the Arg-150 side chain and is selected for representation.

The results of our electrostatic calculations are shown using the 3.5 μs snapshot from the Pdx1-DNA(CTAAT) (*insulin*-like; Figure 2-8A) and Pdx1-DNA(TTAAT) (*iapp*-like; Figure 2-8B) trajectories in Figure 2-8. Given that the entire surface of the DNA duplex bears a strong negative charge, mapping of electrostatic charge onto the molecular surface has been colorized such that the majority of the DNA appears in white, with only those regions that are anomalously negative in charge trending towards deep red. Visualizing the surface charge of the DNA in this way immediately leads to the prediction that binding to the 5′-CTAAT-3′ (*insulin*-like) sequences is enthalpically favorable over binding to 5′-TTAAT-3′ (*iapp*-like) sequences. This prediction is consistent with our thermodynamic data, which show a steep and unfavorable change in the binding enthalpy of *iapp*-derived sequences, as compared to those derived from the *insulin* promoter or the consensus sequence. The deep red pocket occupied by Arg-150 in Figure 2-8A spans both Thy-1 and the 5′ C·G pair. Throughout the trajectory, Arg-150 preferentially orients within this pocket, while simultaneously maintaining the bridging hydrogen bond geometry previously discussed. In contrast, replacement of the 5′ C·G pair with a 5′ T·A pair interposes the partial positive charge on the six membered ring of the Ade nucleobase, resulting in a collapse of the strongly negative pocket to only include the first base pair of the core TAAT motif. Although hydrogen bonding to the Ade-N3 in a bridging conformation reminiscent of that observed with a 5′-CTAAT-3′ (*insulin*-like) sequence is sterically possible, the side chain of Arg-150 does not form this interaction. Instead, Arg-150 remains oriented towards the strongly negative pocket primarily including the first base pair of the TAAT motif.

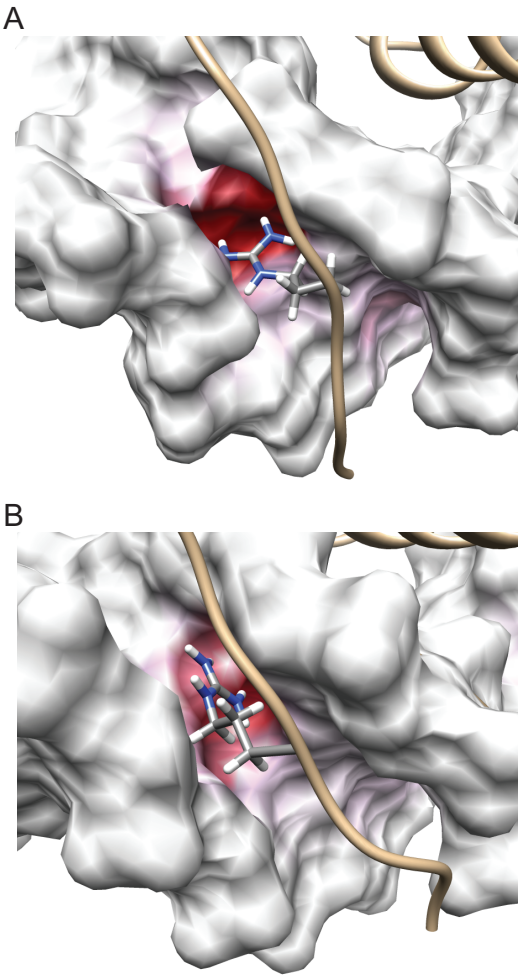

Figure **2-8**. The Arg-150 side chain δ-guanidinium group inserts into a strongly negative patch within the minor groove of the DNA duplex and reorients to track the position of maximum negative charge. The surface electrostatic potential of the DNA, calculated with Pdx1 removed, is mapped onto snapshots from Anton MD trajectories of (A) the Pdx1-DNA_{CON} complex and (B) the Pdx1-DNA(TTAATG) complex. In both cases, the snapshot recorded after 3.5 μs of production dynamics is selected for representation. In order to emphasize the strong negative charge of the Arg-150 binding pocket, even compared with the general charge on a DNA duplex, the color scale of the DNA surface for both panels is -14 kT/e (red) to 0.0 kT/e (white). The side chain atoms of Arg-150 are represented in CPK mode, and colored by atom type (grey, carbon; blue, nitrogen; white, hydrogen).

2.3 Conclusions

Among its many functions, Pdx1 maintains glucose homeostasis by regulating transcription of both the *insulin* and *iapp* genes, through interactions with A-box elements in their promoters. Collectively, the data presented here provide a clear hypothesis for the molecular origins of A-box specificity inherent to the Pdx1-DNA binding mode. The binding specificity of Pdx1 for regulated promoter elements is dominated by the presence of a TAAT sequence, which alone does not provide enough information to impart specificity. Our thermodynamic results quantify Pdx1's preference for Cyt in the 5′-position, relative to the TAAT core, which is fundamental to its differential affinity for the *insulin* and *iapp* promoters. Analysis of microsecond timescale MD simulations, which have been vetted through their ability to reproduce experimental NMR data, reveals a plausible mechanism for the observed sequence-specificity. Our study demonstrates that Pdx1 is able to reject A·T and G·C base pairs at the 5′ position, relative to the TAAT core, because interactions between Arg-150 and the minor groove of the DNA become unstable in this context. More significantly, our study demonstrates that Pdx1 is able to discriminate between *insulin* and *iapp* promoter-derived sequences based on the electrostatic potential created by juxtaposition of their respective 5′-flanking residues with the first T·A pair of the core TAAT sequence, due to preferential electrostatic stabilization of the highly conserved Arg-150 in an orientation that favors hydrogen bonding with the 5′-flanking pair when sequences similar to those found in the human *insulin* promoter are bound.

2.4 Materials and Methods

2.4.1 Protein Preparation and Purification.

A synthetic Pdx1 gene was purchased from Geneart and the Pdx1 homeodomain (amino acids 146-206 of the human sequence; subsequently referred to as Pdx1-HD), Pdx1-HD R150A and Pdx1-HD R150K were subcloned by PCR into pET47b (Novagen) encoding a 6x His tag and a 3C protease recognition site upstream of the cloning site. The recombinant plasmid was transformed into BL21(DE3) competent cells for protein over-expression. Pdx1 constructs for ITC were harvested from cells grown in LB media at 37 °C until an OD of 0.8-1.0 was reached, at which time expression was induced with a final IPTG concentration of 0.5 mM and the temperature reduced to 30 °C. Pdx1-HD for NMR spectroscopy was harvested from cells grown in M9 minimal media, made with $^{15}NH_4Cl$ and either ^{13}C-glucose for backbone assignment or ^{12}C-glucose for spin relaxation measurements. The cells were incubated at 37 °C until an OD of 0.5-0.7 was reached and expression was induced with a final IPTG concentration of 0.5 mM and the temperature reduced to 30 °C. In all cases, cells were harvested 4 hours post-induction and lysed by sonication at 4 °C. The suspension was centrifuged at 4 °C for 30 min at 11500 rpm in a Beckman Coulter Allegra 25R using a TA-14-50 rotor. The supernatant was passed over a Ni-NTA (Novagen) column, and the protein was eluted with imidazole (200 mM). The 6x His-tag was cleaved by incubating with 3C protease at 4°C overnight while also dialyzing away the imidazole. The contents of the dialysis bag were passed over the same NTA-Ni column, and the flow-through was collected. The protein was concentrated using an Amicon Ultra centrifugal filter device (Millipore) that contained a PES 3000 MWCO membrane. The homeodomain was buffer exchanged into the storage buffer (100 mM cacodylate, pH 7.3, and 100 mM potassium chloride). Protein concentrations were determined by UV absorption at 278 nm in denaturing

conditions (final concentration of 6.1 M Guanidinium Chloride, pH 7.5) using the extinction coefficient of 14,000 M^{-1} cm^{-1}.

2.4.2 Duplex DNA Preparation for ITC

All DNA was purchased from Integrated DNA Technologies, Inc (for sequences of the forward-strand see Tables 2-1 and 2-2; all oligonucleotides were purchased with sequences providing perfect Watson-Crick complementarities). ssDNA concentrations were determined based on the IDT calculated extinction coefficients. Duplex formation was accomplished by heating complementary ssDNA in a water bath and slowly cooling to room temperature. Duplex DNA concentrations were quantified by standard UV absorbance assay at 260 nm. Finally, duplex DNA was dialyzed overnight in storage buffer.

2.4.3 Isothermal Titration Calorimetry

All binding studies were performed at 25 °C using standard protocols,[46] adapted for running on an Auto-ITC200 (MicroCal, Inc.). Prior to ITC, duplex DNA and Pdx1-HD were co-dialyzed overnight in storage buffer. 1.2 – 1.5 μL of Pdx1 homeodomain (60 or 100 μM) were injected into 16-mer duplex DNA (6 or 10 μM) at 220 second intervals while the contents of the cell were stirred at a speed of 1000 rpm. The heat of ligand dilution was accounted for by averaging the final five points of each titration and subtracting the average from each point in the titration. ITC experiments for each duplex were completed at the least in triplicate and the averages of the best-fit parameters from a single-site binding model are reported with their associated uncertainties represented by the standard error of the mean. All data was fit in Origin 7.0 (MicroCal, Inc.).

Titrations with Pdx1-HD mutants (R150A and R150K) were performed on a VP-ITC (MicroCal, Inc.) at 25 °C. Samples were prepared for ITC as described above for Pdx1-HD titrations. However, 10 μL of Pdx1-HD mutants (120 μM) was titrated into 16-mer duplex DNA (10-12 μM) with 360 seconds spacing while the contents were stirred at 360 rpm. The data was analyzed and fit as described above for Pdx1-HD.

2.4.4 Anton Simulations of Pdx1

Starting coordinates for all Pdx1 trajectories were derived from the crystal structure of the Pdx1 homeodomain bound to duplex DNA (pdb 2h1k; b, e, and f chains used).[22] The DNA in this starting structure is similar to the consensus duplex used in our experimental studies except for the presence of a single-nucleotide 5′ overhang on both strands; for the wild-type DNA trajectory, the sequence of the DNA simulated was 5′-TCTCTAATGAGTTTC-3′ in complex with 5′-AGAAACTCATTAGAG-3′. The apo-Pdx1 simulation was initiated using the Pdx1 coordinates from the holo-state crystal structure, but with the DNA eliminated; Pdx1-DNA(WT) was initiated directly from the co-crystal structure. Simulations with substituted Watson-Crick base pairs in the 5′-position relative to the TAAT core recognition site were generated from the wild-type co-crystal structure using the Amber nucleic acid builder.[47] Systematic variation of the base pair composition at the 5′-position in the wild type DNA(CTAATG) sequence resulted in three DNA-mutant trajectories, denoted throughout the text as Pdx1-DNA(ATAATG), Pdx1-DNA(GTAATG), and Pdx1-DNA(TTAATG).

All-atom molecular simulations of apo-Pdx1 and Pdx1 in complex with duplex DNA of varying composition were conducted using the ff99SB Amber force field[48] on Anton,[33] a special-purpose machine for molecular mechanics. For all simulations, the starting conformation of Pdx1 or the Pdx1-DNA complex was embedded in a cubic box of SPC water molecules with a bulk

sodium chloride concentration of 100 mM. For apo-Pdx1, this resulted in an initial box roughly 62 Å to a side, containing 14 Na^+ and 23 Cl^- ions (to neutralize the system), and 7,449 water molecules. Each of the four Pdx1-DNA complexes was embedded in a box that began approximately 77.4 Å to a side, containing 46 Na^+ and 27 Cl^- ions, and approximately 14,500 water molecules. The systems were prepared in Maestro for energy minimization and temperature equilibration in an NVT ensemble for a total of 1.2 ns in Desmond (D.E. Shaw Research). Subsequently, each simulation was equilibrated in the NPT ensemble (300 K, 1 bar; Berendsen coupling scheme) for a minimum of 5 ns. All bond lengths to hydrogen atoms were constrained using SHAKE. Production dynamics were run under the same NPT conditions with SHAKE, using a 2.0 fs timestep (long-range electrostatics computed every 6.0 fs). Each simulation was run for a minimum of 5.0 μs of production dynamics. Prior to analysis, snapshots that were stored to disk every 200 ps were stripped of waters using Visual Molecular Dynamics (VMD),[49] which was also used for preliminary trajectory visualization.

2.4.5 Molecular Dynamics Analysis

Molecular graphics images were created using the UCSF Chimera package.[50] Additional analysis and visualization were accomplished in MATLAB (MathWorks). MD-derived order parameters were obtained by using iRED analysis[43] of MD trajectories averaged over 5 ns windows, as previously reported.[51] Backbone carbon chemical shifts were calculated from the MD ensemble using the SHIFTS program,[52] following the protocol of Li and Brüschweiler.[53]

2.4.6 Nonlinear Poisson-Boltzmann Analysis

Electrostatic potential calculations were carried out using numerical solutions to the nonlinear Poisson-Boltzmann equation, as implemented in the adaptive Poisson-Boltzmann solver (APBS)[54] plug-in to PyMOL.[55] Settings were chosen as described by Veeraraghavan *et al*.;[56] briefly, the DNA from each analyzed snapshot was placed in a medium with a dielectric constant of 2.0 within the solvent-accessible surface-enclosed volume and dielectric constant 80 within the continuum solvent containing 0.15M monovalent ions.

2.4.7 Nuclear Magnetic Resonance Spectroscopy

Standard triple resonance NMR techniques were used to assign all resonances of Pdx1-HD (apo) and DNA bound Pdx1-HD (holo) on a BrukerAvance III 500 and 850 MHz spectrometers.[57] Both spectrometers were equipped with TCI cryroprobes for maximum sensitivity. Spectra were processed by NMRpipe and analyzed with SPARKY (SPARKY 3.113; T. D. Goddard and D. G. Kneller, University of California, San Francisco, CA). Relaxation data were analyzed in Matlab (MathWorks). Backbone chemical shifts for both states were assigned from spectra collected at 298K and have been deposited in the BMRB (ID number 19227 for apo-Pdx1 and 19228 for the Pdx1-DNA_{CON} complex).

Experiments measuring ^{15}N T_1 and T_2 spin relaxation, as well as the ^{1}H-^{15}N NOE, were performed at 298 K on a Bruker Avance III 500 MHz spectrometer using standard ^{15}N relaxation methods.[36; 37] Relaxation decays of T_1 were generated by acquiring a set of ten spectra with delays, 50, 150, 150, 250, 350, 450, 550, 550, 650, and 750 ms. T_2 relaxation decays were obtained by using 16, 48, 48, 80, 112, 144, 176, and 208 ms delays. Heteronuclear NOE

experiments were acquired in duplicate using a 5 second saturation period for NOE transfer (and a matching time delay in the control experiment).

2.4.8 Model Free Analysis

Lipari-Szabo modelfree analysis was performed using modelfree4.20,[58] with diffusion tensor fitting using the quadric method[59; 60] and in-house written Matlab scripts as previously described.[61] The x-ray crystal structure of Pdx1-HD (pdb 2h1k)[22] was modified to remove the DNA and used as the structural reference for axially symmetric diffusion tensor estimation. The ^{15}N T_1, T_2, and ^{1}H-^{15}N NOE data were fit using the axially symmetric global diffusion parameters established through the quadric protocol, with S^2 and τ_e (model 2) describing internal motions for each site.

2.5 Acknowledgements

We thank Dr. Neela Yennawar and Julia Fecko for assistance with the Auto-ITC200, and Dr. Debashish Sahu for helpful discussions of homeodomain-DNA interactions and function. This work was supported by an NIH predoctoral fellowship to M.B. (F31GM101936) and by start-up funds provided to S.A.S. by the Pennsylvania State University. The Anton machine at NRBSC/PSC was generously made available by D.E. Shaw Research. Anton computer time was provided by the National Resource for Biomedical Supercomputing (NRBSC) and the Pittsburgh Supercomputing Center (PSC) through Grant RC2GM093307 from the National Institutes of Health.

2.6 References

1. Sander, M. & German, M. S. (1997). The beta cell transcription factors and development of the pancreas. *J Mol Med* **75**, 327-40.

2. Hutton, J. C., Penn, E. J. & Peshavaria, M. (1982). Isolation and characterisation of insulin secretory granules from a rat islet cell tumour. *Diabetologia* **23**, 365-73.

3. Suckale, J. & Solimena, M. (2010). The insulin secretory granule as a signaling hub. *Trends Endocrinol Metab* **21**, 599-609.

4. Ogawa, A., Harris, V., McCorkle, S. K., Unger, R. H. & Luskey, K. L. (1990). Amylin secretion from the rat pancreas and its selective loss after streptozotocin treatment. *J Clin Invest* **85**, 973-6.

5. Westermark, P., Andersson, A. & Westermark, G. T. (2011). Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. *Physiol Rev* **91**, 795-826.

6. Lutz, T. A. (2012). Control of energy homeostasis by amylin. *Cell Mol Life Sci* **69**, 1947-65.

7. Rushing, P. A., Hagan, M. M., Seeley, R. J., Lutz, T. A. & Woods, S. C. (2000). Amylin: a novel action in the brain to reduce body weight. *Endocrinology* **141**, 850-3.

8. Lutz, T. A. (2006). Amylinergic control of food intake. *Physiol Behav* **89**, 465-71.

9. Woerle, H. J., Albrecht, M., Linke, R., Zschau, S., Neumann, C., Nicolaus, M., Gerich, J. E., Goke, B. & Schirra, J. (2008). Impaired hyperglycemia-induced delay in gastric emptying in patients with type 1 diabetes deficient for islet amyloid polypeptide. *Diabetes Care* **31**, 2325-31.

10. German, M. S., Moss, L. G., Wang, J. & Rutter, W. J. (1992). The insulin and islet amyloid polypeptide genes contain similar cell-specific promoter elements that bind identical beta-cell nuclear complexes. *Mol Cell Biol* **12**, 1777-88.

11. German, M., Ashcroft, S., Docherty, K., Edlund, H., Edlund, T., Goodison, S., Imura, H., Kennedy, G., Madsen, O., Melloul, D. & et al. (1995). The insulin gene promoter. A simplified nomenclature. *Diabetes* **44**, 1002-4.

12. Bernardo, A. S., Hay, C. W. & Docherty, K. (2008). Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic beta cell. *Mol Cell Endocrinol* **294**, 1-9.

13. Glick, E., Leshkowitz, D. & Walker, M. D. (2000). Transcription factor BETA2 acts cooperatively with E2A and PDX1 to activate the insulin gene promoter. *J Biol Chem* **275**, 2199-204.

14. Cissell, M. A., Zhao, L., Sussel, L., Henderson, E. & Stein, R. (2003). Transcription factor occupancy of the insulin gene in vivo. Evidence for direct regulation by Nkx2.2. *J Biol Chem* **278**, 751-6.

15. Ohneda, K., Mirmira, R. G., Wang, J., Johnson, J. D. & German, M. S. (2000). The homeodomain of PDX-1 mediates multiple protein-protein interactions in the formation of a transcriptional activation complex on the insulin promoter. *Mol Cell Biol* **20**, 900-11.

16. Serup, P., Jensen, J., Andersen, F. G., Jorgensen, M. C., Blume, N., Holst, J. J. & Madsen, O. D. (1996). Induction of insulin and islet amyloid polypeptide production in pancreatic islet glucagonoma cells by insulin promoter factor 1. *Proc Natl Acad Sci U S A* **93**, 9015-20.

17. Chakrabarti, S. K., James, J. C. & Mirmira, R. G. (2002). Quantitative assessment of gene targeting in vitro and in vivo by the pancreatic transcription factor, Pdx1. Importance of chromatin structure in directing promoter binding. *J Biol Chem* **277**, 13286-93.

18. Miller, C. P., McGehee, R. E., Jr. & Habener, J. F. (1994). IDX-1: a new homeodomain transcription factor expressed in rat pancreatic islets and duodenum that transactivates the somatostatin gene. *EMBO J* **13**, 1145-56.

19. Liberzon, A., Ridner, G. & Walker, M. D. (2004). Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1. *Nucleic Acids Res* **32**, 54-64.

20. Hay, C. W. & Docherty, K. (2006). Comparative analysis of insulin gene promoters: implications for diabetes research. *Diabetes* **55**, 3201-13.

21. Babu, D. A., Deering, T. G. & Mirmira, R. G. (2007). A feat of metabolic proportions: Pdx1 orchestrates islet development and function in the maintenance of glucose homeostasis. *Mol Genet Metab* **92**, 43-55.

22. Longo, A., Guanga, G. P. & Rose, R. B. (2007). Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain. *Biochemistry* **46**, 2948-57.

23. German, M. S. & Wang, J. (1994). The insulin gene contains multiple transcriptional elements that respond to glucose. *Mol Cell Biol* **14**, 4067-75.

24. Macfarlane, W. M., Campbell, S. C., Elrick, L. J., Oates, V., Bermano, G., Lindley, K. J., Aynsley-Green, A., Dunne, M. J., James, R. F. & Docherty, K. (2000). Glucose regulates islet amyloid polypeptide gene transcription in a PDX1- and calcium-dependent manner. *J Biol Chem* **275**, 15330-5.

25. Rudnick, A., Ling, T. Y., Odagiri, H., Rutter, W. J. & German, M. S. (1994). Pancreatic beta cells express a diverse set of homeobox genes. *Proc Natl Acad Sci U S A* **91**, 12203-7.

26. MacFarlane, W. M., Read, M. L., Gilligan, M., Bujalska, I. & Docherty, K. (1994). Glucose modulates the binding activity of the beta-cell transcription factor IUF1 in a phosphorylation-dependent manner. *Biochem J* **303 (Pt 2)**, 625-31.

27. Taylor, D. G., Babu, D. & Mirmira, R. G. (2005). The C-terminal domain of the beta cell homeodomain factor Nkx6.1 enhances sequence-selective DNA binding at the insulin promoter. *Biochemistry* **44**, 11269-78.

28. Francis, J., Babu, D. A., Deering, T. G., Chakrabarti, S. K., Garmey, J. C., Evans-Molina, C., Taylor, D. G. & Mirmira, R. G. (2006). Role of chromatin accessibility in the occupancy and transcription of the insulin gene by the pancreatic and duodenal homeobox factor 1. *Mol Endocrinol* **20**, 3133-45.

29. Ades, S. E. & Sauer, R. T. (1995). Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry* **34**, 14601-8.

30. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry* **38**, 1999-2017.

31. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* **19**, 120-7.

32. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu Rev Biophys* **41**, 429-52.

33. Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J., Chao, J. C., Eastwood, M. P., Gagliardo, J., Grossman, J. P., Ho, C. R., Ierardi, D. J., Kolossvary, I., Klepeis, J. L., Layman, T., Mcleavey, C., Moraes, M. A., Mueller, R., Priest, E. C., Shan, Y. B., Spengler, J., Theobald, M., Towles, B. & Wang, S. C. (2008). Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the Acm* **51**, 91-97.

34. Billeter, M., Qian, Y. Q., Otting, G., Muller, M., Gehring, W. & Wuthrich, K. (1993). Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. *J Mol Biol* **234**, 1084-93.

35. Marsh, J. A., Singh, V. K., Jia, Z. & Forman-Kay, J. D. (2006). Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci* **15**, 2795-804.

36. Palmer, A. G., 3rd. (2004). NMR characterization of the dynamics of biomacromolecules. *Chem Rev* **104**, 3623-40.

37. Jarymowycz, V. A. & Stone, M. J. (2006). Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences. *Chem Rev* **106**, 1624-71.

38. Bruschweiler, R. (2003). New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Current Opinion in Structural Biology* **13**, 175-183.

39. Lipari, G. & Szabo, A. (1982). Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity. *Journal of the American Chemical Society* **104**, 4546-4559.

40. Anderson, J. S., Forman, M. D., Modleski, S., Dahlquist, F. W. & Baxter, S. M. (2000). Cooperative ordering in homeodomain-DNA recognition: solution structure and dynamics of the MATa1 homeodomain. *Biochemistry* **39**, 10045-54.

41. Fausti, S., Weiler, S., Cuniberti, C., Hwang, K. J., No, K. T., Gruschus, J. M., Perico, A., Nirenberg, M. & Ferretti, J. A. (2001). Backbone dynamics for the wild type and a double H52R/T56W mutant of the vnd/NK-2 homeodomain from Drosophila melanogaster. *Biochemistry* **40**, 12004-12.

42. Doerdelmann, T., Kojetin, D. J., Baird-Titus, J. M., Solt, L. A., Burris, T. P. & Rance, M. (2012). Structural and biophysical insights into the ligand-free Pitx2 homeodomain and a ring dermoid of the cornea inducing homeodomain mutant. *Biochemistry* **51**, 665-76.

43. Prompers, J. J. & Brüschweiler, R. (2002). General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation. *Journal of the American Chemical Society* **124**, 4522-4534.

44. German, M. S., Blanar, M. A., Nelson, C., Moss, L. G. & Rutter, W. J. (1991). Two related helix-loop-helix proteins participate in separate cell-specific complexes that bind the insulin enhancer. *Mol Endocrinol* **5**, 292-9.

45. Ohlsson, H., Karlsson, K. & Edlund, T. (1993). IPF1, a homeodomain-containing transactivator of the insulin gene. *EMBO J* **12**, 4251-9.

46. Buurma, N. J. & Haq, I. (2007). Advances in the analysis of isothermal titration calorimetry data for ligand-DNA interactions. *Methods* **42**, 162-72.

47. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668-1688.

48. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins-Structure Function and Bioinformatics* **65**, 712-725.

49. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33-38.

50. Pettersen, E. F. G., T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. (2004). UCSF Chimera a visualization system for exploratory research and analysis. *J Comput Chem.* **25**, 1605-1612.

51. Showalter, S. A. & Brüschweiler, R. (2007). Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *Journal of Chemical Theory and Computation* **3**, 961-975.

52. Xu, X. P. & Case, D. A. (2001). Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13 ' chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR* **21**, 321-333.

53. Li, D. W. & Brüschweiler, R. (2010). Certification of Molecular Dynamics Trajectories with NMR Chemical Shifts. *Journal of Physical Chemistry Letters* **1**, 246-248.

54. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* **98**, 10037-41.

55. Schrodinger, LLC. (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

56. Veeraraghavan, N., Ganguly, A., Chen, J. H., Bevilacqua, P. C., Hammes-Schiffer, S. & Golden, B. L. (2011). Metal binding motif in the active site of the HDV ribozyme binds divalent and monovalent ions. *Biochemistry* **50**, 2672-82.

57. Kanelis, V., Forman-Kay, J. D. & Kay, L. E. (2001). Multidimensional NMR methods for protein structure determination. *IUBMB Life* **52**, 291-302.

58. Mandel, A. M., Akke, M. & Palmer, A. G., 3rd. (1995). Backbone dynamics of Escherichia coli ribonuclease HI: correlations with structure and function in an active enzyme. *J Mol Biol* **246**, 144-63.

59. Brüschweiler, R., Liao, X. B. & Wright, P. E. (1995). Long-Range Motional Restrictions in a Multidomain Zinc-Finger Protein from Anisotropic Tumbling. *Science* **268**, 886-889.

60. Lee, L. K., Rance, M., Chazin, W. J. & Palmer, A. G., 3rd. (1997). Rotational diffusion anisotropy of proteins from simultaneous analysis of 15N and 13C alpha nuclear spin relaxation. *J Biomol NMR* **9**, 287-98.

61. Wostenberg, C., Quarles, K. A. & Showalter, S. A. (2010). Dynamic Origins of Differential RNA Binding Function in Two dsRBDs from the miRNA "Microprocessor" Complex. *Biochemistry* **49**, 10728-10736.

Chapter 3

A Short C-terminal Extension of the Pdx1 Homeodomain Modulates dsDNA Affinity

Abstract

Homeobox proteins are vital for anatomical development of higher order species. Their DNA binding motif has been structurally and functionally characterized to a great extent; however, the homeodomain often only makes up a small portion of the polypeptide chain, with regions of disorder spanning the N-terminus, C-terminus or both. In stark contrast to the homeodomains themselves, the disordered regions of homeodomain proteins are not as well characterized nor are their functional roles. In chapter 2, DNA binding studies with the Pdx1 homeodomain resulted in several significant findings. First, Pdx1-HD preferentially interacted with Insulin promoter elements in comparison to IAPP elements, which resulted from stronger electrostatic interactions with Arg-150. Mutational analysis of Pdx1-HD Arg-150 to Ala abrogated binding and mutation to Lys diminished binding affinity and enthalpy. Additionally, single base pair mutations upstream relative to the core binding site demonstrated preferential binding by Pdx1-HD relative to the consensus sequence. These preferences could not be explained. In this chapter, a primarily thermodynamic view of the impact of a short region of the disordered C-terminus on homeodomain function is presented. Additionally, two basic residues potentially involved in specificity are mutated to determine their functional role. Surprisingly, we show that the five C-terminal residues added contributed to the elongation and stabilization of helix 3, likely resulting in the observed increase in stability of the protein. Through isothermal titration calorimetry, we

show that addition of part of the C-terminal domain of Pdx1 enhances the binding affinity for short dsDNA sequences. Overall, this study demonstrates the ability of residues outside the homeodomain to tune binding affinity for preferential binding to their cognate sites.

3.1 Introduction

Many sequence specific interactions arise from the insertion of an α-helix into the major groove of double stranded DNA (dsDNA), enabling nucleotide readout. For example, homeobox proteins contain a dsDNA binding motif that cooperatively folds into the canonical all α-helical domain, in which two parallel helices lie perpendicularly across the recognition helix (helix 3). The homeodomain is a great model system, enabling the study of sequence specificity by major groove and minor groove interactions.[1] Two critical regions of the protein interact with DNA: the disordered N-terminal arm and the recognition helix. The N-terminal arm inserts into the minor groove, while the recognition helix enters the major groove, making additional sequence specific interactions with the 5´-TAAT-3´ core binding site. As a guide for the remainder of this discussion, Figure 3-1 draws attention to the numbering of nucleotides and polarity of the dsDNA encompassing the TAAT core binding site, with respect to the transcription start site. Much of the sequence specificity comes from interactions with the core binding site; however, homeodomains have been shown to recognize sequence in the 3´-flanking region of the core.[2-4] In the presence of a large pool of DNA sequence, these proteins must be able to identify and bind with high affinity to their cognate sites, in addition to discriminating closely related sites.

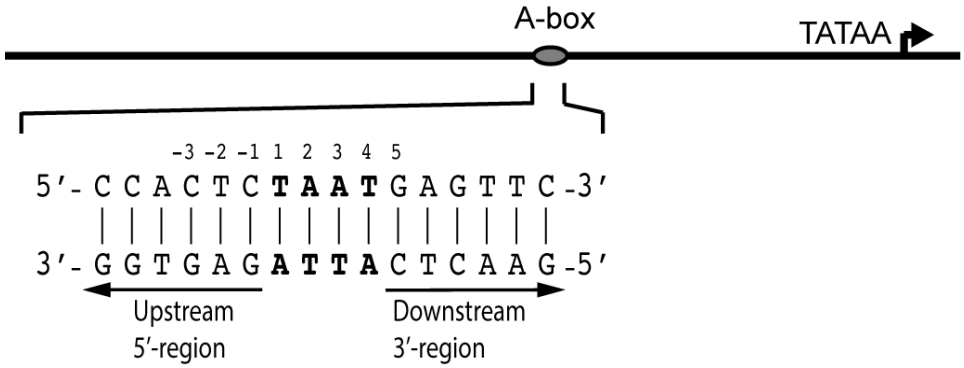

Figure **3-1.** Depiction of Pdx1 binding elements with respect to the transcription start site. The 5´-region refers to nucleotides upstream the core binding site and the 3´-region refers to bases downstream the core binding site. Numbering scheme above the duplex DNA sequence is with respect to the first base in the 5´-TAAT-3´ sequence referred to as 1. Numbering is sequential downstream, whereas negative values depict bases upstream the core site.

Several studies of homeodomain proteins have reported DNA specificity being directed by core flanking nucleotides.[3-5] Of these studies, two carried out a study in which the nucleotide identity was systematically varied.[3, 4] Catron et *al.* studied the DNA binding specificity of not one, but three different homeodomain proteins, ultimately showing that DNA mutations to the 5´-region modestly effect binding, whereas mutations to the 3´-region exhibit more pronounced effects on binding.[3] Furthermore, the systematic DNA mutations effected DNA binding of the homeodomains to a varying degree, consistent with differences in sequence selectivity based on homeodomain sequence. In another study, Ades et *al.* studied the DNA binding affinity of the Engrailed homeodomain protein against systematically mutated DNA sequences within the TAAT core and two bases into 3´-region.[4] The binding affinity of Engrailed was reduced when substitutions to the DNA sequence were made, with an increasing effect from the minor groove to major groove, which is consistent with the notion that more information is stored within the Watson-Crick face in the major groove.[4]

Although binding studies by several groups show that homeodomains discriminate DNA sequence beyond the canonical tetranucleotide core site, the protein residues involved in these

interactions often remain undocumented and general rules are elusive. Co-crystal structures show primarily DNA contacts made by the well-ordered homeodomain, yet the homeodomain is only a portion of the protein. Additional interactions involving non-homeodomain amino acids may contribute to sequence specificity of the homeodomain. Because variation in the amino acid sequence within the recognition helix alters the sequence selectivity of the homeodomain, it is plausible that the sequence of the regions outside the homeodomain influence the sequence selectivity of the homeodomain. In turn, this may provide the additional sequence selectivity necessary to discriminate cognate sites from non-cognate binding sites.

Many homeodomain proteins have regions of disorder at the N-terminus, C-terminus, or both. An excellent example of this is the homeodomain protein Pdx1, which has a long disordered C-terminal tail in addition to a disordered N-terminal domain. Pdx1 activates transcription of various peptide hormones by binding to AT-rich A-boxes located in the promoter region of these hormone genes.[6-10] In particular *insulin* and *islet amyloid polypeptide (IAPP)* are transcriptionally activated by Pdx1. Interestingly, partial sequence alignment of Pdx1 (homeodomain and C-terminus) demonstrates a significant conservation of the homeodomain, but less conservation of the disordered C-terminal tail across species. Importantly, a positively charged patch proximal to the homeodomain is well conserved and highlighted in Figure 3-2. To better understand the role of the C-terminal tail with respect to the homeodomain, a construct referred to as Pdx1-HDC encompassing both domains was generated. Unfortunately, Pdx1-HDC was not stable under any of the many buffer conditions tested. Following purification of Pdx1-HDC, degradation products were typically observed within 48 hours. Mass spectrometry was employed to determine the region where degradation was catalyzed. As a result, a shorter construct encompassing a basic patch proximal to the homeodomain was generated (Pdx1-HDx) to study the effects on homeodomain DNA binding.

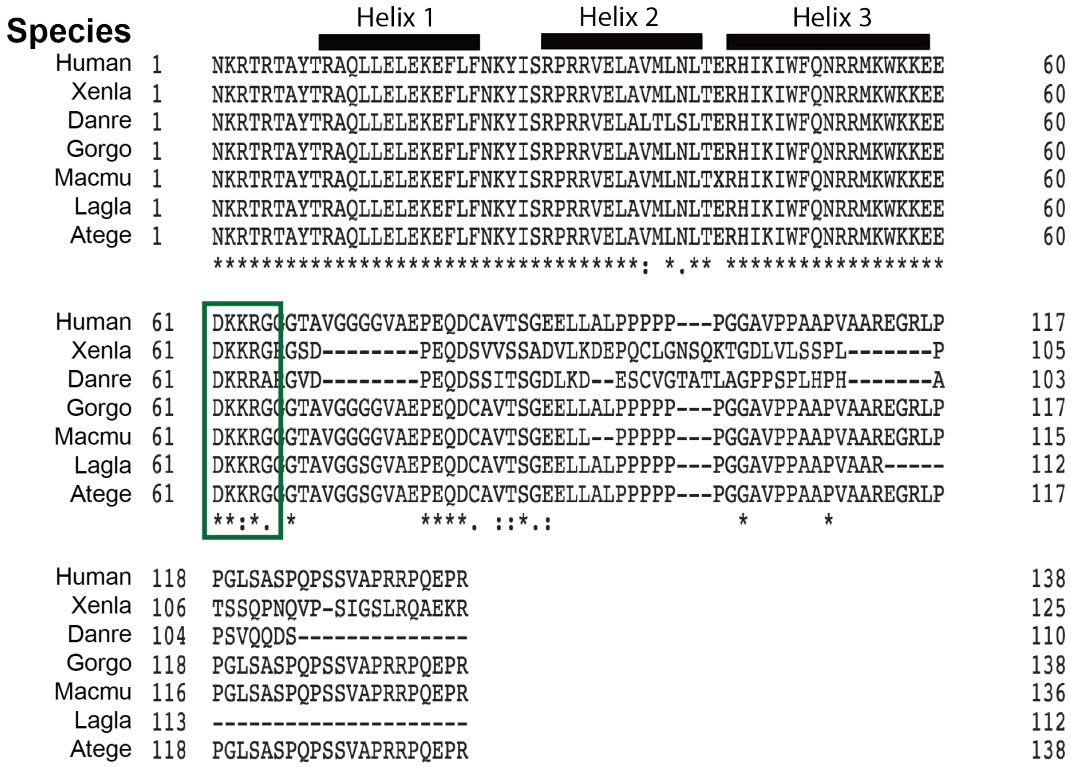

Figure **3-2.** Sequence alignment of Pdx1 homeodomain through the end of the polypeptide sequence for several species: Human, *Xenopus leavis* (Xenla), Zebrafish (Danre), Gorilla (Gorgo), Macaque (Macmu), Woolly monkey (Lagla), and Spider monkey (Atege). The secondary structure elements are annotated above the figure. The box highlights the conserved basic patch. The Pdx1-HDx construct investigated in this chapter encompasses residues 1-65 in this figure, which translates to residues 146-211 in the full-length human Pdx1 sequence.

In this study, we investigate the impact on DNA binding by Pdx1-HDx and examine the DNA binding effects of DNA point mutations in the flanking regions. Moreover, in an effort to identify additional residues important for imparting specificity, DNA binding by two single mutant proteins was investigated. Using ITC, we can learn how the additional basic residues impact DNA binding affinity and specificity. By NMR methods, structural information of the protein can be assessed to determine impact on structure by the additional residues in comparison to the homeodomain-only construct. We report on the ability of Pdx1-HDx to discriminate

sequence at the 5´-region of the core binding site and enhance the overall DNA binding affinity of the homeodomain.

3.2 Materials and Methods

3.2.1 Protein Preparation and Purification

Pdx1 HDx (amino acids 146-211) was generated for ITC and NMR using methods described previously.[11] Briefly, the gene was subcloned into pET47b+ between *EcoRI* and *XmaI* cut sites. The recombinant plasmid was transformed into BL21 (DE3) cells, grown at 37 °C until an OD of 0.8-1.0 was reached. The cells were induced with 0.5 mM IPTG and allowed to overexpress at 30 °C for 4 hours. Similarly, NMR samples were prepared and purified as previously described.[11] Protein concentration was determined by UV absorbance at 278 nm using the extinction coefficient of 14,000 M^{-1} cm^{-1}. The extinction coefficient of Pdx1 HDx was theoretically determined using the Scripps Protein Calculator program (http://protcalc.sourceforge.net/).

3.2.2 DNA Preparation

Complementary single stranded DNA oligomers were purchased from Integrated DNA Technologies, Inc. The single stranded DNA concentrations were determined using the molar extinction coefficients provided by Integrated DNA Technologies, Inc on the specification sheets. Samples were reconstituted in 100 mM Cacodylate, 100 mM KCl, pH 7.3. Equimolar amounts of complementary single stranded oligomers were heated in a water bath to 95 °C and allowed to

cool slowly to room temperature. Duplex DNA concentrations were quantified by UV absorbance at 260 nm. Duplexed samples were co-dialyzed overnight with Pdx1 HDx in separate membranes.

3.2.3 Isothermal Titration Calorimetry

All titrations were completed on an Auto-iTC 200 calorimeter (MicroCal, Inc.) equipped with a robotic housing. Titrations were set up as described previously with the cell containing the dsDNA (10-12 μM) and the syringe containing Pdx1-HDx (120-125 μM).[11] Both protein and DNA sequences were co-dialyzed in 100 mM cacodylate, 100 mM KCl, pH 7.3. Data was processed using MicroCal Origin Software (10.2.1) for ITC-200. Prior to fitting the data using a one site binding model, a baseline correction was made by averaging the last five points of the upper baseline. Because the Insulin E1 titration resulted in weak binding, as expected, a protein into buffer titration was subtracted from the Insulin E1 titrations prior to data fitting.

3.2.4 Nuclear Magnetic Resonance Spectroscopy

Concentrations of Pdx1-HDx for NMR spectroscopy were approximately 500 μM and 1.0 mM for relaxation experiments and backbone assignments, respectively. The protein was buffer exchanged into 100 mM Cacodylate, pH 6.5, 100 mM KCl, 10% (w/v), and 0.01% NaN_3. For relaxation studies, the buffer was 50 mM Cacodylate, pH 6.5 (final pH), 50 mM KCl, 5 mM TCEP, 10% (w/v), and 0.01% NaN_3. Standard proton-detect triple resonance experiments were used to assign the backbone. All experiments were collected at 298K on a BruckerAvance II 500 MHz spectrometer equipped with TCI cryoprobes. For data analysis, spectra were first processed by NMRPipe and analyzed in SPARKY (SPARKY 3.113; T. D. Goddard and D. G. Kneller,

University of California, San Francisco, CA). Spin relaxation data were analyzed in Matlab (MathWorks).

^{15}N spin T_1, and T_2 relaxation and NOE experiments were performed on a Bruker Avance III 500 MHz spectrometer using standard ^{15}N relaxation methods.[12, 13] Relaxation decays of T_1 were generated by acquiring a set of ten spectra with delays, 50, 150, 250, 350, 450, 550, 650, and 750 ms. T_2 relaxation decays were obtained by using 16, 48, 80, 112, 144, 176, and 208 ms delays. The $^{1}H,^{15}N$-NOE data set was generated by acquiring the spectra with and without NOE.

3.2.5 Anton Simulations of Pdx1

All Pdx1 trajectories were obtained from the crystal structure of Pdx1 (pdb 2h1k, chains b, e, and f).[14] Similar to the experimental consensus sequence, the 16 base pair DNA sequence used in the simulations was 5′-TCTCTAATGAGTTTC-3′ and the complementary strand was 5′-AGAAACTCATTAGAG-3′. The complex was started directly from the co-crystal structure. As previously described, molecular simulations were performed using the ff99SB Amber force field[15] on Anton[16].[11]

3.2.6 Molecular Dynamics Analysis

Images were generated using the UCSF Chimera package.[17] Further analysis was completed in MATLAB (MathWorks).

3.3 Results and Discussion

3.3.1 Chemical Shift Analysis and Conformational Dynamics of Pdx1-HDx

A complete set of backbone chemical shifts of Pdx1-HDx was assigned by acquiring standard double and triple resonance experiments on a 500 MHz spectrometer. Chemical shift assignments were then used to calculate the secondary structure propensity (SSP) of the protein to verify proper folding of Pdx1-HDx and monitor variations in secondary structure of the extension, if any. The SSP program was used to calculate ΔδCα-ΔδCβ, which reports on secondary structural motifs by comparing the experimental data set to a reference data set (chemical shifts of proteins with known structures).[18] Positive variation from the reference values indicates α-helix, whereas negative values represent β-structure or coil structure. As illustrated in Figure 3-3, SSP analysis yields three distinct α-helices represented by significant positive values (≥ +2.0). As a guide, the secondary structural elements from the co-crystal structure of Pdx1 are annotated above the figure. Values less than zero represent turns or β-structure (≤ -2.0); for Pdx1, we expect a loop and a turn between helix 1 and 2, and helix 2 and 3, respectively. Furthermore, values near zero, such as those observed for residues 146 to 152 in the N-terminal arm are disordered. Comparing the SSP analysis of Pdx1-HDx to the homeodomain-only (Figure 2-4) displays significant stability of the helices.[11] We find that helix 3 becomes stabilized by the addition of the 5 residues from the C-terminus, as evidenced by the SSP values ≥ +2.0 up to residue 206. This is in comparison with the homeodomain-only results (Figure 2-4) for which helix 3 ends at residues 196. This finding is consistent with the increase in stability of the Pdx1-HDx construct from that of Pdx1- HD observed when concentrating the proteins. In order to independently document the extent to which a stable helix is formed in this region, we next turned to spin relaxation data.

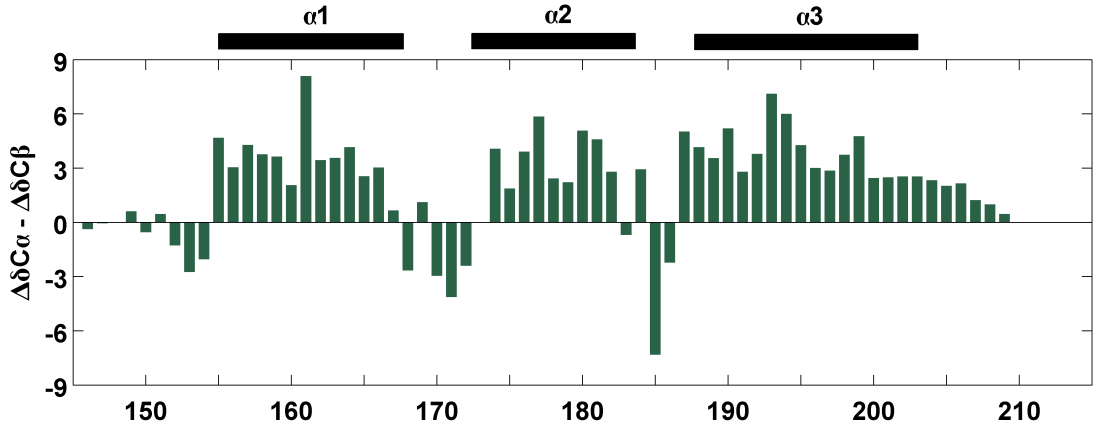

Figure **3-3.** Chemical shift analysis using secondary C^{α} and C^{β} chemical shifts representing the secondary structure of Pdx1-HDx. Positive stretches greater than 2 indicates stable α-helical structure, whereas negative stretches indicate turns or β-structure. The helices corresponding to the co-crystal structure are annotated above for referencing. The x-axis respresents the amino acid residue in the Pdx1 sequence.

Spin relaxation provides quantification of protein backbone motions on the picosecond to nanosecond time scales as measured by ^{15}N T_1, T_2, and ^{1}H-^{15}N NOE.[12,13,19] For disordered regions, such as the N-terminal arm and the far C-terminus of Pdx1-HDx, we expect the dynamics of these regions to be enhanced. The data in Figure 3-4 show an increase in the motions of the turn and the loop regions with a slight increase in T_1 and T_2 values and a slight decrease in NOE values. A more pronounced enhancement in the relaxation of the N-terminal arm and the C-terminal region (aa 204-211) is observed. Notably, residues 204-207 in the C-terminus exhibit enhanced dynamics, though not to the extent of residues at the extreme termini, which display greater picosecond to nanosecond dynamics, which is expected due to end effects. This data strongly suggests that helix 3 extends to residue 207, but helix fraying may play a role in increased dynamics.

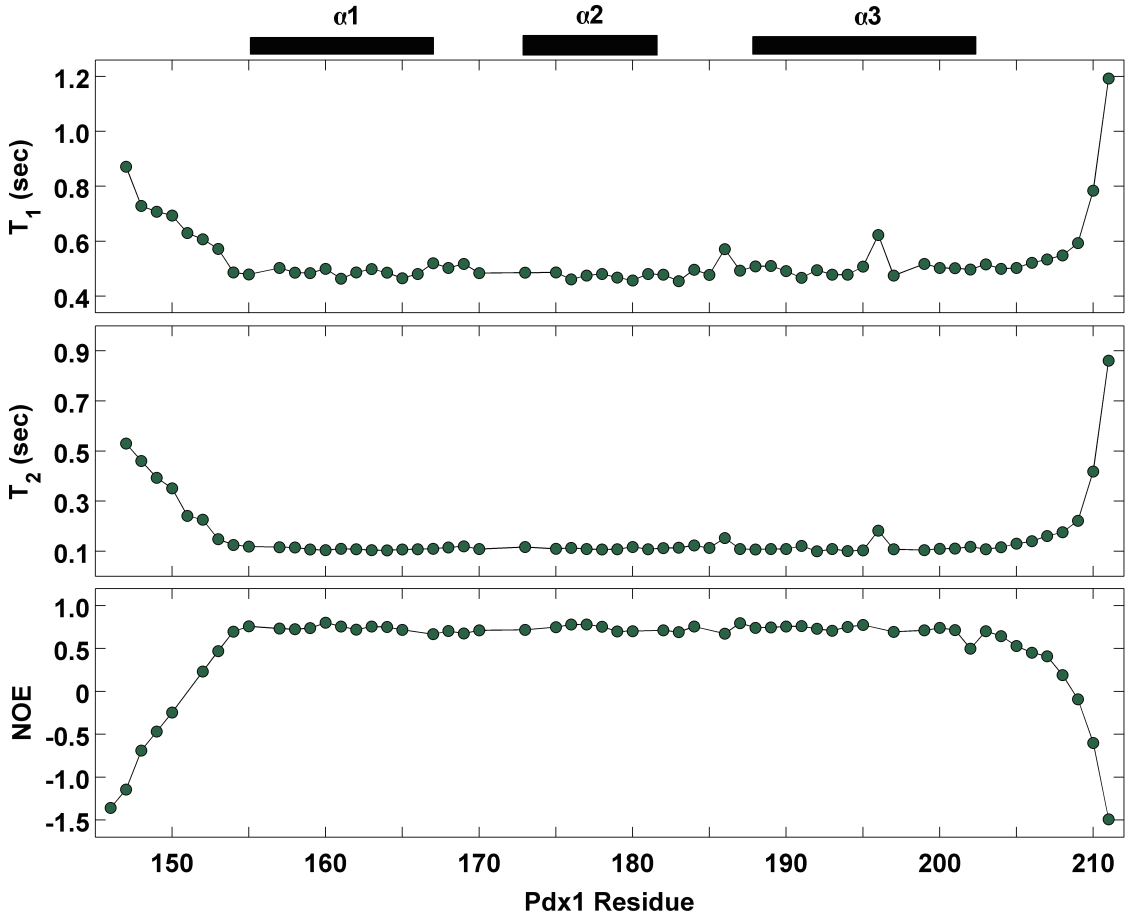

Figure **3-4.** Backbone ^{15}N T_1, T_2, and $^{1}H,^{15}N$-NOE of Pdx1-HDx measured on a 500 MHz NMR spectrometer. Backbone dynamics of Pdx1-HDx reported by (A) ^{15}N-T_1, (B) ^{15}N-T_2, and (C) $^{1}H,^{15}N$-NOE. The secondary structure from the co-crystal structure is annotated above the figure.

For well-folded proteins, the baseline in the T_1 and T_2 data is an indicator of protein size. For larger proteins, the T_1 baseline will increase and the T_2 baseline will decrease if the magnetic field strength of the measurement is held constant. This is to say, if a protein undergoes oligomerization, one can identify this occurrence by qualitatively analyzing the baselines and comparing the overall correlation time, τ_c, to a theoretically computed value. For Pdx1-HDx, the average T_1 relaxation for HDx is about 0.5 seconds, whereas for the disordered or highly flexible tails the relaxation time is much higher. As expected for a highly mobile amide, the T_1 value would be larger because rapid motions are slower to relax. A similar trend for the T_2 relaxation is

observed. However, the average baseline for Pdx-HDx for T_2 relaxation is near 0.11 ms, with greater values for the highly flexible tails. For a typical homeodomain, with data collected on an 11.7 T magnet, the T_2 value is on the order of 0.15 seconds, where here the value is closer to 0.1 seconds, which is expected due to the slight increase in size of Pdx1-HDx compared to the canonical homeodomain.[11, 20] In summary, for a protein of this particular size, the T_1 and T_2 values on a per residue basis are reasonable.

To analyze the oligomerization state of Pdx1-HDx, an estimation of the overall rotational correlation time (τ_c) was determined using two independent methods; the Stokes-Einstein-Debye (SED) equation and the R_2/R_1 (or T_1/T_2) estimation initially reported by Kay and colleagues.[19] Computing τ_c using the SED theory yields a analytical value under the assumption that the volume of the protein is spherical. For a given protein as a function of temperature, τ_c is given by equation 3-1[21],

$$\boldsymbol{\tau_c = \frac{4\pi\eta}{3k_BT} r^3} \tag{3.1}$$

where k_B is the Boltzmann constant, T is the temperature in Kelvin,

$$\boldsymbol{\eta = 1.7753 - (0.0565T_C) + (1.0751\times10^{-3}T_C^2) - (9.2222\times10^{-6}T_C^3)} \tag{3.2}$$

is the viscosity[22] (η) as a function of T_C (Celsius), and

$$\boldsymbol{r = \sqrt[3]{\frac{3vM}{4\pi N_A}} + r_w} \tag{3.3}$$

is the estimated hydrodynamic radius of a protein of a given molecular weight. In Equation 3-3[21], v is the partial specific volume of 0.73 cm^3/g, M is the molecular weight in Daltons, and N_A is Avogadro's number and r_w is the radius of hydration set to 3.2 Å.

Using the equation 3-1, τ_c was computed based on the molecular weight of the ^{15}N,^{13}C-labeled protein (8950 Da), resulting in a value of 4.41 ns. As reported by Kay and colleagues, the R2/R1 ratio can be used to estimate the rotational correlation of a protein.[19] The correlation time

obtained from the baseline average R_2/R_1 ratio for Pdx1-HDx was 4.24 ns. Both values are in good agreement strongly suggesting the tumbling properties of Pdx1-HDx are isotropic in solution for a monomeric oligomerization state.

For well established secondary structure, such as for the three α-helices forming the homeodomain, rather large positive NOEs are observed. The NOE values of Pdx1-HDx display a highly rigid backbone for the majority of the protein, with an average NOE of 0.74, which is near the theoretical maximum of ~0.82 (at 500 MHz).[19] Increased flexibility is observed for the N-terminal arm, with some values falling below zero indicative of extreme flexibility. The C-terminal extension displays marked flexibility as evidenced by the decrease in NOE values from that of the α-helices. Molecular dynamic simulations model a short α-helix in the C-terminal extension; the NOE values certainly reflect a more stable structure extending up to residue 207, although residues 205-207 reflect a slight increase in dynamics compared to the α-helices. This decrease in rigidity may be due to helical fraying. Regions with increased flexibility yield NOEs that fall well below the theoretical maximum; in some cases negative NOES are observed for extremely flexible regions such as the N-terminal arm and C-terminal residues of Pdx1-HDx. Residues 207-211, in particular, result in slightly decreased NOEs, which maybe be due to partial helical content due to fraying of the recognition helix.

3.3.2 Investigating the Interaction of Pdx1-HDx with the Consensus DNA Sequence by Molecular Dynamics Simulations

We turned to molecular dynamics simulations to investigate the structure and function of the C-terminal extension with regards to the homeodomain. During the five microsecond trajectory, the C-terminal extension trails over an entire face of one side of the DNA, interacting with bases in the minor grove, DNA backbone, and perhaps weak interactions with bases in the

major groove. These interactions can be easier visualized in the contact map shown in Figure 3-5, which illustrates the protein-protein and protein-DNA contacts. The four tiles represent intra-protein interactions, protein-DNA contacts and DNA-DNA interactions. The top left quadrant and bottom right quadrant are equivalent and represent the protein-DNA contacts. Intra-protein contacts are represented in the top right panel. Anti-parallel DNA-DNA interactions are observed in the bottom left quadrant. The diagonal shown, in the top right quadrant, in white is artificially set to zero as far as i to i+3 interactions so that these strong, but uninformative intra-protein contacts do not bleach out the color scale, whereas residues making contact are set to the value proportional to the frequency in the five microsecond trajectory. A clear interaction between Lys147 or Arg150 in the N-terminal arm and the DNA can be observed. Furthermore, several residues in the N-terminal arm interact with the complementary strand. Interestingly, a weak interaction with Arg155 is observed with the far 5´-region of the minor groove (visualized in Figure 3-5), leading to the hypothesis that Arg155 imparts sequence selectivity upstream relative to the core binding site.

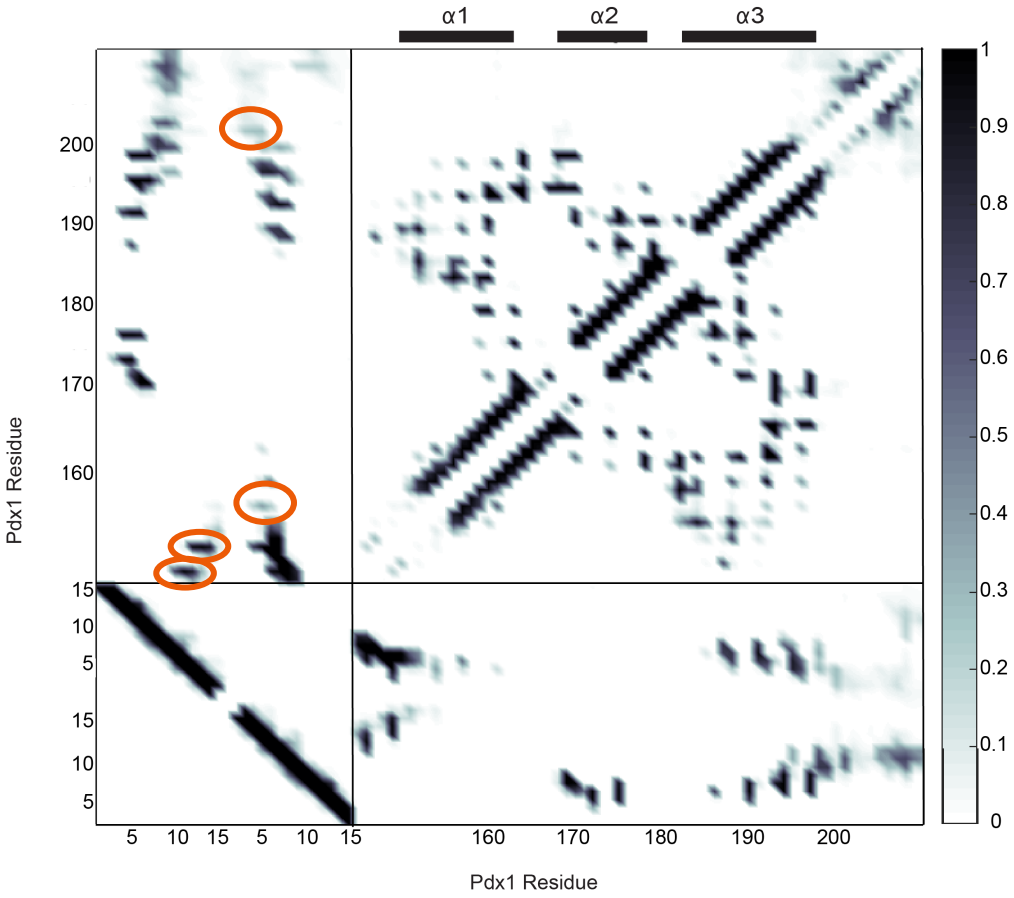

Figure **3-5.** A contact map representation for the co-crystal structure of Pdx1-HDx with the consensus sequence. The diagonals are set to zero, whereas neighbors ≤ 0.6 nm maintaining contact are set to a value that is proportional to the frequency of the observed contact. Lys147, Arg150, Arg155, and Lys203 are circled in orange.

Additionally, two basic patches (aa 202-203 and aa 207-209) in the C-terminus of Pdx1-HDx interact with the DNA; these interactions primarily occur in the 3´-region of the binding site. On the opposing strand, residues from these two basic patches interact to a lesser extent in the minor groove. More specifically, Lys203 interacts moderately with the DNA in the 5´-region upstream that core site. Finally, the molecular dynamics simulations led us to hypothesize that basic residues Arg155 and Lys203 provide additional sequence specificity. In order to test this

hypothesis, we generated the corresponding mutants in our experimental construct to investigate their binding to DNA in parallel with the wild-type Pdx1-HDx study discussed below.

3.3.3 Binding of Pdx1-HDx to Promoter-Derived Sequences

The binding of Pdx1-HDx to elements from the human Insulin and IAPP promoters was studied thermodynamically using ITC. The calorimetric data was fit using a one-site binding model to yield the thermodynamic parameters: n, K_a, and ΔH. Subsequently, the Gibbs free energy (ΔG) of the reaction was derived using the following equation

$$\Delta \boldsymbol{G} = -\boldsymbol{RTln}(\boldsymbol{K_a}) \tag{3.4}$$

where R is the universal gas constant in kcal K^{-1} mol^{-1}, T is the temperature of the experiment in Kelvin, and K_a is the binding constant in M^{-1}. Furthermore, the entropy of binding (ΔS) for the system is calculated using

$$-\boldsymbol{T}\Delta \boldsymbol{S} = \Delta \boldsymbol{G} - \Delta \boldsymbol{H} \tag{3.5}$$

where T is the temperature in Kelvin, and ΔH is the binding enthalpy from the fitted data. Table 3-1 shows the thermodynamic data describing the binding of Pdx1-HDx to a panel of promoter-derived sequences.

Table **3-1**. Thermodynamic parameters describing binding of Pdx1-HDx to dsDNA derived from natural human promoter sequences measured at 298K. All error represents standard error of the mean estimated from at least three replicate measurements with fitting performed in Origin 10.2.1 (MicroCal, Inc.).

DNA	**Forward Sequence**	**n**	**K_d (nM)**	**ΔH (kcal/mol)**	**ΔG (kcal/mol)**	**-TΔS (kcal/mol)**
Insulin A1	5′-AGGCCC**TAAT**GGGCCA-3′	0.8 ± 0.1	4.9 ± 0.5	-8.80 ± 0.02	-11.4 ± 0.1	-2.6± 0.1
Insulin A3	5′-AGACTC**TAAT**GACCCG-3′	0.9 ± 0.1	6 ± 1	-9.03 ± 0.03	-11.3 ± 0.1	-2.3 ± 0.1
Insulin E1	5′-AGCCATCTGCCGACCC-3′	1.0 ± 0.1	280 ± 30	-2.17 ± 0.03	-9.0 ± 0.1	-6.8 ± 0.1
IAPP A1	5′-GGAAAT**TAAT**GACAGA-3′	1.0 ± 0.1	14 ± 2	-7.49 ± 0.05	-11.1 ± 0.1	-3.6 ± 0.1
IAPP A2	5′-ATGAGT**TAAT**GTAATA-3′	1.1 ± 0.1	23 ± 5	-4.89 ± 0.05	-10.4 ± 0.1	-5.6 ± 0.1
IAPP A1 Mut	5′-GGA**C**AT**TAAT**GACAGA-3′	1.1 ± 0.1	2 ± 1	-7.41 ± 0.07	-11.8 ± 0.3	-4.4 ± 0.3

First, as a control for non-sequence specific binding by Pdx1-HDx, binding to the non-cognate Insulin E1 sequence was measured. A relatively weak binding affinity of 280 ± 30 nM and low binding enthalpy of -2.17 ± 0.03 kcal/mol was observed for Pdx1-HDx, consistent with non-specific interactions involving the repressor MetJ and its cognate DNA.[23, 24] Moreover, the entropic contribution for non-specific binding is dominant. For Pdx1-HDx binding to Insulin E1, a large entropic contribution is computed, -TΔS = -6.76 kcal/mol, which is consistent with the expected non-specific binding by Pdx1-HDx to this sequence. Interestingly, a similar pattern for ΔH and –TΔS is observed for the IAPP A2 sequence. Pdx1-HDx binds with a significantly weaker affinity for Insulin E1, as expected for a non-cognate binding site. Although IAPP A2 contains the core binding site, binding by Pdx1-HDx displays a more favorable entropic contribution, similar to what was observed for the non-specific binding to Insulin E1. These thermodynamic contributions differ from the determined contributions of Insulin A1, Insulin A3, and IAPP A1 (Table 3-1).

Pdx1-HDx bound to elements from the Insulin promoter tighter than elements from the IAPP promoter, following a similar trend as previously reported with the homeodomain-only construct from Chapter 2. The extended Pdx1-HDx construct favored binding to Insulin promoter

elements, as evidenced by tighter binding and more favorable binding enthalpies. Specifically, the Pdx1-HDx binding affinities for Insulin elements were near 5 nM and the binding enthalpies were approximately -9 kcal/mol. This trend was also observed with the homeodomain-only construct described in Chapter 2. Binding to IAPP elements resulted in an approximate 3-fold lower binding affinity and the binding enthalpies were less favorable (~6 kcal/mol) than those for Insulin elements. The trend observed for Pdx1-HDx promoter binding parallels the trend reported for the homeodomain-only construct, except a slight decrease from the 5-fold difference between elements is observed. Nevertheless, all promoter-derived binding events translated into large negative ΔG values resulting in ΔG of ca. -11 kcal/mol for Pdx1-HDx sequence-specific interactions and ΔG of ca. -9 kcal/mol for Insulin E1 (non-sequence specific). The magnitude and differences between sequence and non-sequence specific interactions are consistent with the literature.[24, 25]

Interestingly, the IAPP sequences have much higher AT content in the 5´-region relative to the core and overall in comparison to Insulin and Consensus sequences. The trinucleotide sequence upstream the core site is similar in all three positions in the IAPP promoter elements. In position -1 from the core, the IAPP promoter elements contain a Thy and similarly, in position -3 from the core, the sequences contain an Ade. In fact, the C(-3)A mutation results in a decreased binding affinity from 9 nM for the consensus sequence to 17 nM (Table 3-2). In common with the IAPP promoter elements, this trinucleotide site contains an Ade at the -3 position, which perhaps plays a major role in the binding preference of Pdx1-HDx.

To test the hypothesis that the nucleotide identity in position -3 plays a role in sequence selectivity of Pdx1-HDx, a mutant IAPP A1 sequence was generated. The mutant sequence incorporated the consensus nucleotide in position -3; Ade was mutated to Cyt. The expectation is that this mutated sequence would restore a tighter binding affinity, similar to that of the consensus sequence. Indeed, the interaction between the IAPP A1 mutant sequence and Pdx1-HDx was 2

nM as compared with the affinity of 14 nM for IAPP A1. Interestingly, the enthalpy remained the same with a ΔΔH of 0.08 kcal mol^{-1}, suggesting the mechanism of binding is similar. Based on the co-crystal structure, there are no major contacts made with DNA upstream from the core site; however, the thermodynamics presented in Table 3-1 demonstrate otherwise. An investigation into the amino acid(s) responsible for imparting specificity in position -3 was completed.

Overall, Pdx1-HDx binding to promoter-derived elements was similar to that reported for the homeodomain-only construct in Chapter 2. Unique to this construct, the binding to the non-specific DNA element, Insulin E1 was much tighter (~4.5-fold) than was determined for Pdx1-HD. Preferential binding was still observed for the Pdx1-HDx construct, which suggests the additional basic residues in this construct likely contribute to binding affinity and not selectivity. To determine the effect of the additional residues in binding selectivity in the flanking regions of the DNA, ITC experiments were collected with a panel of dsDNA sequence derived from the consensus sequence.

3.3.4 Altered Thermodynamics of Pdx1 Homeodomain Extended Construct

A detailed binding study of Pdx1-HDx with a large panel of DNA sequences derived from the consensus sequence was carried out using ITC. The calorimetric parameters, stoichiometry (n), enthalpy (ΔH), and binding affinity (K_d) for these sequences are tabulated (Table 3-2). Derived from a sequence alignment of Pdx1's putative binding elements from several promoter regions, Liberzon and colleagues identified a consensus sequence.[26] Together with the sequence used in the co-crystal structure of Pdx1 homeodomain, the Liberzon consensus sequence was slightly modified to yield a longer consensus sequence, which formed the basis of all of our derived sequences shown in Table 3-2.[14, 26] Under equivalent sample and buffer conditions as all experiments, a second event (i.e. conformational transition, change in

oligomerization state, or binding) occurred for Cons C(-1)A resulting in an erroneous binding affinity and was excluded from the panel. Using the consensus sequence as the baseline for comparison, we identified sequences in which the affinity of Pdx1-HDx increased or decreased approximately two-fold.

Table **3-2.** Thermodynamic parameters describing binding of Pdx1-HDx to dsDNA derived from consensus recognition sequences measured at 298K. All error represents standard error of the mean estimated from at least three replicate measurements with fitting performed in Origin 10.2.1 (MicroCal, Inc.).

DNA	**Forward Sequence**	**n**	**K_d (nM)**	**ΔH (kcal/mol)**
Consensus	5′-CCACTC**TAAT**GAGTTC-3′	1.1 ± 0.1	9 ± 1	-8.94 ± 0.04
Cons 5′C(-1)G	5′- CCACTG**TAAT**GAGTTC -3′	1.1 ± 0.1	2.1 ± 0.5	-8.60 ± 0.05
Cons 5′C(-1)T	5′- CCACTT**TAAT**GAGTTC -3′	0.9 ± 0.1	2 ± 1	-7.68 ± 0.05
Cons 5′T(-2)G	5′- CCACG**CTAAT**GAGTTC -3′	1.1 ± 0.1	3 ± 1	-10.53 ± 0.05
Cons 5′T(-2)C	5′- CCACC**CTAAT**GAGTTC -3′	1.1 ± 0.1	1.4 ± 0.3	-10.53 ± 0.04
Cons 5′T(-2)A	5′- CCACA**CTAAT**GAGTTC -3′	0.9 ± 0.1	9 ± 1	-9.9 ± 0.1
Cons 5′C(-3)G	5′- CCAG**TCTAAT**GAGTTC -3′	1.1 ± 0.1	4 ± 1	-8.68 ± 0.04
Cons 5′C(-3)T	5′- CCAT**TCTAAT**GAGTTC -3′	1.1 ± 0.1	7 ± 2	-7.82 ± 0.03
Cons 5′C(-3)A	5′- CCAA**TCTAAT**GAGTTC -3′	1.1 ± 0.1	17 ± 3	-7.13 ± 0.05
Cons 3′G(5)T	5′-CCACTC**TAATT**AGTTC-3′	0.9 ± 0.1	0.37 ± 0.07	-9.12 ± 0.04
Cons 3′G(5)A	5′-CCACTC**TAATA**AGTTC-3′	1.2 ± 0.1	6 ± 2	-3.51 ± 0.05
Cons 3′G(5)C	5′-CCACTC**TAATC**AGTTC-3′	1.1 ± 0.1	8 ± 1	-9.02 ± 0.04

To determine the effects of DNA sequence in the flanking regions on DNA binding, we investigated the affinities and binding enthalpies of Pdx1-HDx with each of the consensus mutants. For 5 of the derived sequences designed to monitor Pdx1-HDx binding affinity upstream from the core binding site, binding was tighter in comparison to the consensus. However, three sequences did not result in increased affinity. In particular, the sequences T(-2)A, C(-3)T, C(-3)A,

resulted in an equivalent affinity to consensus or 2-fold weaker affinity. This finding of weak effect suggests that the homeodomain interacts with flanking bases, including bases in the minor groove where the N-terminal arm interacts. Notably, the data demonstrates that Pdx1-HDx is preferential to the sequence upstream the core binding site as observed in the variation in binding affinities, especially up to 3 base pairs from the core binding site.

Pdx1-HDx interacted more strongly with sequences having higher GC content, in particular those in which an AT (or TA) base pair was mutated to CG (or GC). Although Cyt was mutated to Thy, the Cons C(-1)T mutant sequence demonstrated 4.5-fold favorable binding affinity by Pdx1, but the binding enthalpy was less favorable than that of consensus. Binding to either Cons C(-1)G, Cons T(-2)G, and Cons T(-2)C was around 2 nM, which was 4-fold stronger than consensus; the binding enthalpies were similar or more favorable than that of the consensus sequence. Overall, the results in Table 3-2 show that mutating a base pair in the 5´-region to an AT (or TA) weakens the binding affinity of Pdx1-HDx and results in a slight increase the binding enthalpy.

Comparing binding affinities of modifications made upstream or downstream from the core results in clear differences in affinities. Focusing on the base substitutions made to the 3´-region, Pdx1-HDx binding was preferential towards a Thy in position 5, consistent with previous binding studies.[11] Conversely, Pdx1-HDx binding to Gua, Cyt, or Ade resulted in similar affinities near 8 nM within error. Interestingly, binding to G(5)A resulted in a drastically weaker binding enthalpy reflective of non-sequence specific binding, yet this sequence contains the core binding site. Thus, it is likely that mutating the wild type Gua to an Ade disrupts sequence specific hydrogen bonding.[27]

The trend in affinities tabulated for Pdx1-HDx parallel the trend observed for the homeodomain-only construct. The only significant difference is the binding affinity, which is consistently tighter for all DNA constructs in this study. Table 3-2 supports the notion that the

binding mode for Pdx1-HDx is consistent with Pdx1-HD and the additional residues impact binding by stabilizing helix 3 or by interactions made with the DNA.

3.3.5 Arg155 Contributes to Electrostatic Interactions with the dsDNA

Molecular Dynamic simulations illuminated the potential role of Arg155 in DNA binding. Arg155 protrudes from the protein near the 5´-region of the minor groove, potentially interacting with bases within the region. Five microsecond simulations revealed a transient interaction with Arg155 and bases in the minor groove, providing evidence for Arg155's role in selectivity upstream that core binding site (Figure 3-5). We hypothesized that Arg155 interacted with bases upstream the core binding site contributing to the sequence selectivity of Pdx1-HDx. Thus, to test this hypothesis, an R155A mutant was generated and screened for binding with several DNA sequences from our pool of sequences.

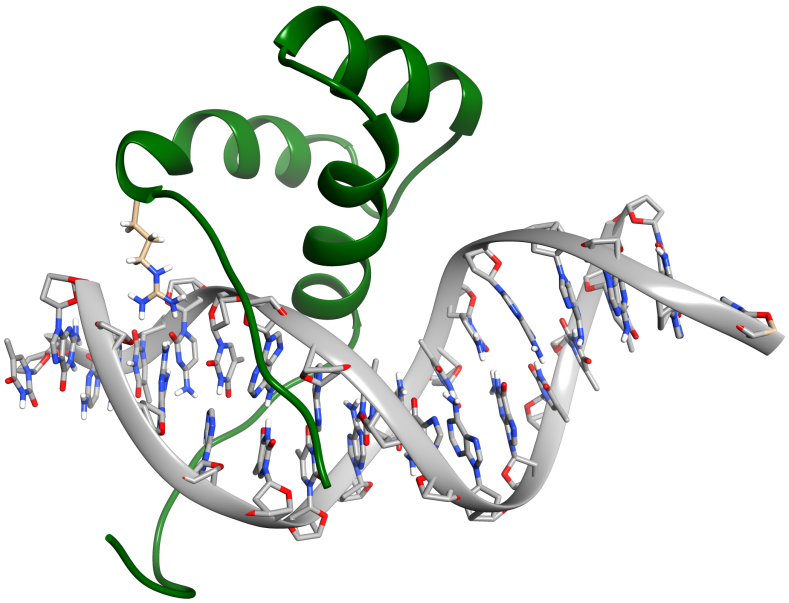

Figure **3-6.** A representative frame from the 5 µs molecular dynamic simulation illustrating Arg-155 interacting in the minor groove. The side chain of Arg-155 is colored in tan, whereas the remainder of Pdx1-HDx is colored green.

Our calorimetric results in Table 3-3 show that R155A only slightly modulates the binding affinity of a select few sequences we screened, including consensus, Insulin A3, and IAPP A1. The binding affinity of R155A for the consensus sequence remained equivalent to Pdx1-HDx. However, the R155A binding affinity for Insulin A3 and IAPP A1 decreased by approximately 2- and 1.5-fold, respectively. No clear trend to explain these results could be identified. Because the changes observed were very small, there are two possibilities to explain these observations. Arg155 does not appreciably contribute to the general DNA binding affinity or the conformers observed in the MD simulations are not representative of the true dynamics of the complex.

The binding enthalpies of Pdx1-HDx R155A and consensus, Insulin A3, or IAPP A1 are approximately 1-2 kcal/mol smaller than enthalpies of Pdx1-HDx for each DNA. For consensus and Insulin A3 sequence, which have similar sequences in the minor groove, resulted in a binding enthalpy reduction of 1.2 kcal/mol. However, for the IAPP A1 sequence, a more significant reduction in the binding enthalpy of 2.3 kcal/mol compared to Pdx1-HDx was observed. Because the binding enthalpies are affected more significantly than the affinities, this supports the notion that Arg155 is involved in electrostatic interactions, perhaps mediated by water molecules.

Table **3-3.** Thermodynamic parameters comparing binding of Pdx1-HDx, Pdx1-HDx R155A, Pdx1-HDx K203A to dsDNA derived from consensus recognition sequences measured at 298K. All error represents standard error of the mean estimated from at least three replicate measurements with fitting performed in Origin 10.2.1 (MicroCal, Inc.).

		HDx		**HDx R155A**		**HDx K203A**	
DNA	**Forward Sequence**	**K_d (nM)**	**ΔH (kcal/mol)**	**K_d (nM)**	**ΔH (kcal/mol)**	**K_d (nM)**	**ΔH (kcal/mol)**
Insulin A3	5′-AGACTC**TAAT**GACCCG-3′	4.9 ± 0.5	-8.80 ± 0.02	13 ± 1	-7.86 ± 0.04	8 ± 1	-8.79 ± 0.01
Consensus	5′-CCACTC**TAAT**GAGTTC-3′	9 ± 1	-8.94 ± 0.04	8 ± 1	-7.67 ± 0.04	6 ±1	-8.7 ± 0.1
IAPP A1	5′-GGAAAT**TAAT**GACAGA-3′	14 ± 2	-7.49 ± 0.05	22 ± 4	-5.12 ± 0.04		

3.3.6 Lysine Residue in Helix 3 Does Not Contribute to Binding Specificity

In an effort to identify the residue(s) responsible for sequence selectivity upstream from the core binding site, we turned to molecular dynamic simulations to identify potential targets for mutagenesis to alanine residues. Two basic residues in C-terminus of helix 3 were potential candidates based on molecular dynamic simulations of the homeodomain-only construct. This was explored by ITC methods with the consensus and Insulin A3 sequences.

Shown in Table 3-3, titrations of the K203A mutant with consensus and Insulin A3 resulted in binding affinities that were comparable to those of Pdx1-HDx. The K203A mutant interacted with consensus and Insulin A3 with an affinity of 6 ± 1 nM and 8 ± 2 nM. Similarly, the binding enthalpies were similar to the enthalpies of Pdx1-HDx, suggesting the mutation did not impact binding. Furthermore, this suggests that Lys203 does not interact with bases in the major groove to contribute to affinity or selectivity. It is likely that other basic residues in the extension enhance the binding affinity by making contacts with the backbone and bases.

3.4 Discussion

3.4.1 Functional Effect of C-terminal Extension and Alanine Mutations

Residues within the homeodomain are highly conserved in mammals as evidenced in the partial sequence alignment in Figure 3-5. In vertebrate Pdx1 sequences the residues in the homeodomain are conserved with near 100% sequence identity. Considering the homeodomain binds DNA in a sequence specific manner, it is not shocking that the homeodomain sequence is strongly conserved in vertebrates. However, strikingly, a small patch of positively charged

residues proximal to the C-terminus of the homeodomain is well conserved in the Pdx1 sequence, which is highlighted in Figure 3-5. This sequence, which is DKKR may tune the affinity of the homeodomain by interactions with DNA. Our thermodynamic results reported in this study demonstrate an enhancement in binding affinity by the homeodomain, with the addition of DKKRRGG from the mammalian Pdx1 sequence. These charged residues tightened the binding affinity of the homeodomain likely contributed by electrostatic interactions with the DNA. The calorimetric data demonstrates sequence selectivity by the homeodomain, however further studies to identify the residue(s) responsible for said sequence selectivity will illuminate the mechanism.

3.4.2 Structural Characteristics of Pdx1-HDx

Chemical shift analysis and measurement of ^{15}N backbone relaxations provides structural information for proteins, while spin relaxation also provides information on backbone motions occurring on the picosecond to nanosecond timescale. Motions on these timescales can be correlated to functional roles such as protein-protein interactions and molecular recognition of DNA by proteins. For Pdx1-HDx, although the relative homeodomain fold remained primarily constant, slight variations in the secondary structure were observed at the C-terminus of the homeodomain. Addition of the five residue extension stabilized helix 3 resulting in a more stable protein altogether. On the other hand, the dynamics exhibited by the N-terminal arm and the extension residues (KKRGG) matched that of disordered residues, which we show play a role in modulating DNA binding.

3.5 Contribution

Dr. Scott A. Showalter performed and processed molecular dynamic simulations of Pdx1-HDx with the consensus sequence.

3.6 References

1. Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure, *Cell 131*, 530-543.

2. Ekker, S. C., Young, K. E., von Kessler, D. P., and Beachy, P. A. (1991) Optimal DNA sequence recognition by the Ultrabithorax homeodomain of Drosophila, *The EMBO journal 10*, 1179-1186.

3. Catron, K. M., Iler, N., and Abate, C. (1993) Nucleotides flanking a conserved TAAT core dictate the DNA binding specificity of three murine homeodomain proteins, *Molecular and cellular biology 13*, 2354-2365.

4. Ades, S. E., and Sauer, R. T. (1995) Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex, *Biochemistry 34*, 14601-14608.

5. Liu, Y., Matthews, K. S., and Bondos, S. E. (2008) Multiple intrinsically disordered sequences alter DNA binding by the homeodomain of the Drosophila hox protein ultrabithorax, *The Journal of biological chemistry 283*, 20874-20887.

6. Boam, D. S., and Docherty, K. (1989) A tissue-specific nuclear factor binds to multiple sites in the human insulin-gene enhancer, *The Biochemical journal 264*, 233-239.

7. German, M. S., Moss, L. G., Wang, J., and Rutter, W. J. (1992) The insulin and islet amyloid polypeptide genes contain similar cell-specific promoter elements that bind identical beta-cell nuclear complexes, *Molecular and cellular biology 12*, 1777-1788.

8. Watada, H., Kajimoto, Y., Umayahara, Y., Matsuoka, T., Kaneto, H., Fujitani, Y., Kamada, T., Kawamori, R., and Yamasaki, Y. (1996) The human glucokinase gene beta-cell-type promoter: an essential role of insulin promoter factor 1/PDX-1 in its activation in HIT-T15 cells, *Diabetes 45*, 1478-1488.

9. Watada, H., Kajimoto, Y., Kaneto, H., Matsuoka, T., Fujitani, Y., Miyazaki, J., and Yamasaki, Y. (1996) Involvement of the homeodomain-containing transcription factor PDX-1 in islet amyloid polypeptide gene transcription, *Biochemical and biophysical research communications 229*, 746-751.

10. Waeber, G., Thompson, N., Nicod, P., and Bonny, C. (1996) Transcriptional activation of the GLUT2 gene by the IPF-1/STF-1/IDX-1 homeobox factor, *Molecular endocrinology 10*, 1327-1334.

11. Bastidas, M., and Showalter, S. A. (2013) Thermodynamic and structural determinants of differential Pdx1 binding to elements from the insulin and IAPP promoters, *Journal of molecular biology 425*, 3360-3377.

12. Jarymowycz, V. A., and Stone, M. J. (2006) Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences, *Chemical reviews 106*, 1624-1671.

13. Palmer, A. G., 3rd. (2004) NMR characterization of the dynamics of biomacromolecules, *Chemical reviews 104*, 3623-3640.

14. Longo, A., Guanga, G. P., and Rose, R. B. (2007) Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain, *Biochemistry 46*, 2948-2957.

15. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters, *Proteins Struct. Funct. Bioinf. 65*, 712-725.

16. Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J., Chao, J. C., Eastwood, M. P., Gagliardo, J., Grossman, J. P., Ho, C. R., Ierardi, D. J., Kolossvary, I., Klepeis, J. L., Layman, T., Mcleavey, C., Moraes, M. A., Mueller, R., Priest, E. C., Shan, Y. B., Spengler, J., Theobald, M., Towles, B., and Wang, S. C. (2008) Anton, a special-purpose machine for molecular dynamics simulation, *Commun Acm 51*, 91-97.

17. Pettersen, E. F. G., T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. (2004) UCSF Chimera a visualization system for exploratory research and analysis, *J Comput Chem. 25*, 1605-1612.

18. Marsh, J. A., Singh, V. K., Jia, Z., and Forman-Kay, J. D. (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation, *Protein science : a publication of the Protein Society 15*, 2795-2804.

19. Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease, *Biochemistry 28*, 8972-8979.

20. Doerdelmann, T., Kojetin, D. J., Baird-Titus, J. M., Solt, L. A., Burris, T. P., and Rance, M. (2012) Structural and biophysical insights into the ligand-free Pitx2 homeodomain and a ring dermoid of the cornea inducing homeodomain mutant, *Biochemistry 51*, 665-676.

21. Cavanagh, J., Fairbrother, W. J., Palmer, A. G., 3rd, Skelton, N. J., and Rance, M. (2010) *Protein NMR Spectroscopy: Principles and practice*, 2 ed., Academic Press.

22. Garcia de la Torre, J., Huertas, M. L., and Carrasco, B. (2000) HYDRONMR: Prediction of NMR relaxation of globular proteins from atomic level strucutres and hydrodynamic calculations, *J. Magn. Reson. 147*, 138-146.

23. Cooper, A., McAlpine, A., and Stockley, P. G. (1994) Calorimetric studies of the energetics of protein-DNA interactions in the E. coli methionine repressor (MetJ) system, *FEBS letters 348*, 41-45.

24. Hyre, D. E., and Spicer, L. D. (1995) Thermodynamic evaluation of binding interactions in the methionine repressor system of Escherichia coli using isothermal titration calorimetry, *Biochemistry 34*, 3212-3221.

25. Ladbury, J. E. (1995) Counting the calories to stay in the groove, *Structure 3*, 635-639.

26. Liberzon, A., Ridner, G., and Walker, M. D. (2004) Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1, *Nucleic acids research 32*, 54-64.

27. Ladbury, J. E. (2004) Application of isothermal titration calorimetry in the biological sciences: things are heating up!, *BioTechniques 37*, 885-887.

Chapter 4

A Primer for Carbon-Detected NMR Applications to Intrinsically Disordered Proteins in Solution

[Modified from the published paper titled " A Primer for Carbon-Detected NMR Applications to Intrinsically Disordered Proteins in Solution" by Monique Bastidas, Eric B. Gibbs, Debashish Sahu, Scott A. Showalter in *Concept. Magn. Reson. Part A*, (2015), in press at time of thesis writing. Parallel experiments involving the protein Fcp1 were performed by Eric Gibbs and are not reported in this thesis. Debashish Sahu was responsible for developing the NMR pulse programs used to generate Figure 4.]

Abstract

Characterization of intrinsically disordered proteins (IDPs) has increased tremendously over the past two decades. NMR-based structural characterization has been widely embraced by the IDP community, largely because this technique is amenable to highly flexible biomolecules. Particularly, carbon-detect NMR experiments provide a straightforward and expedient method for completing backbone assignments, thus providing the framework to study the structural and dynamic properties of IDPs. However, these experiments remain unfamiliar to most NMR spectroscopists, thus limiting the breadth of their application. In an effort to remove barriers that may prevent the application of carbon-detected bio-NMR where it has the potential to benefit investigators, here we describe the experimental requirements to collect a robust set of carbon-detected NMR data for complete backbone assignment of IDPs. Specifically, we advocate the use of 3D experiments that exploit magnetization transfer pathways initiated on the aliphatic protons, which produces increased sensitivity and provides a suitable method for IDPs that are only

soluble in basic pH conditions (>7.5). The applicability of this strategy to systems featuring a high degree of proline content will also be discussed.

4.1 Introduction

NMR spectroscopy is the most powerful method available for determining the conformational and dynamic properties of highly flexible biopolymers in solution. Intrinsically disordered proteins (IDPs) have been embraced by the NMR community in recent years because their disordered nature pushes the boundaries of structural biology.[1] Unfortunately, the very same conformational disorder that makes IDPs fascinating to study often leads to poor spectral quality, particularly when traditional proton-detected, triple resonance NMR approaches are utilized. Solving this problem has been a focus of our laboratory for several years. Here we describe a framework to study structure-function relationships for IDPs that relies exclusively on carbon-detected solution NMR spectroscopy.[2] We find this strategy to be generally effective and have applied it to an IDP that is nearly indistinguishable from a random coil.[3]

An excellent general introduction to carbon-detected solution NMR strategies that are applicable to biomolecules was previously published by Bermel et *al*.[4] In their review, Bermel *et al.* expertly address several of the key issues general to carbon-detected measurements on uniformly ^{13}C-enriched biomolecules, placing strong emphasis on the "virtual decoupling" methods that are employed to suppress scalar coupling in the direct-detected dimension.[4] Here we will present several pulse programs that utilize in-phase / anti-phase virtual decoupling and refer the reader to the above mentioned review if this concept is unfamiliar.

Until recently, carbon-detected NMR strategies applicable to proteins in solution were built around so-called "protonless" pulse sequences, in which pulsing on the ^{1}H-nucleus was

scrupulously avoided for both excitation and detection.[5] The choice to remain rigorously "protonless" was motivated by an interest in studying metaloenzymes with paramagnetic centers in their active sites. About five years prior to the publication of this article, we and others began to realize that carbon-detected strategies were also a highly efficient means to study IDPs. Without paramagnetic relaxation enhancement of transverse ^{1}H spin-states present as an experimental design factor, we began incorporating polarization transfer from ^{1}H to lower gyromagnetic ratio heteronuclei in the excitation elements of our pulse programs.[6, 7] At present, virtually all carbon-detected biomolecular NMR experiments have been converted to H-start formats that offer enhanced sensitivity. For IDPs, either care must be taken with these experiments to transfer magnetization from aliphatic protons or, if amide-proton polarization is used to initiate the pulse program, the sample must be kept in mildly acidic conditions in order to avoid overwhelming solvent exchange. Recent advances, including new pulse sequences we introduce here, enable studies of IDPs in basic solution conditions (pH > 7.5). Overall, our experience suggests that carbon-detected NMR should enable the study of virtually any IDP that is of biological interest, with only very modest constraints on the solution conditions needed for experimental feasibility. This review will summarize the best practices we have established for our laboratory in the hope that interested investigators will be able to efficiently utilize this exciting new strategy for their own disordered protein systems.

4.1.1 Experimental Systems used to Provide Case Studies

Our discussion will be motivated through the presentation of spectra collected on the C-terminal domain of the pancreatic and duodenal homeobox protein 1 (Pdx1-C), which is strongly disordered in solution, deficient in acidic or basic amino acid residues, and highly enriched in

proline and glycine residues.[3] Pdx1-C is an excellent model IDP because of its amino acid sequence bias including high proline content, which may deter the pursuit of backbone assignments and its lack of structural propensity. Taken together, Pdx1-C offers a subset of the extreme conditions spectroscopists investigating IDPs may encounter.

4.2 Pulse Program and Spectrometer Selection

The traditional chemical shift strategy has relied on the classical $^{1}H,^{15}N$-HSQC or $^{1}H,^{15}N$-TROSY based triple resonance experiments. These strategies are popular because of the ease in unambiguously assigning individual residues as a result of excellent signal dispersion typically seen for cooperatively folded proteins. Furthermore, higher magnetic field strengths provide additional advantages for the traditional proton-based chemical shift strategy, such as improved spectral resolution, reduction of sample concentration or acceleration of data collection, as a result of increased sensitivity. With advancements in probe technology, sample concentration and data collection speed are enhanced when cryogenically cooled probes are utilized.

In the case of carbon-detected biomolecular NMR, experimental strategies can be built around several 2D-detection platforms. In uniformly ^{13}C-enriched samples, it is likely that the detection-nucleus chosen will have a robust one-bond scalar coupling to multiple spin-1/2 nuclei that are suitable to generate the indirect spectral dimension. In this section, we briefly introduce three of the most popular carbon-detected 2D experiments, placing particular emphasis on the common variants of the extremely powerful $^{15}N,^{13}C$-CON experiment. We close the section with a brief commentary on the performance of CON-based experimental strategies at ultra-high magnetic field strength.

4.2.1 Selecting the right 2D correlation experiment for the job

Chemical shift is an inherently local reporter of structure and chemical environment. Structural elements of a protein, such as secondary and tertiary structure, leads to greater signal dispersion in an $^{1}H,^{15}N$-HSQC due to unique chemical environments of individual amino acids. Unlike their well-folded counterparts, IDPs, typically, have severe signal overlap due to their low sequence complexity and lack of secondary and/or tertiary structure. Although severe signal overlap may be an issue of some IDPs, some IDPs exhibit excellent signal dispersion in a $^{1}H,^{15}N$-HSQC, such is the case for Pdx1-C (Figure 4-1A). Relying on the $^{1}H,^{15}N$-HSQC raises another issue for IDPs. For IDPs that have proline rich regions, the $^{1}H,^{15}N$-HSQC does not provide a robust platform for further studies, due to the lack of an amide proton in proline residues. The sequence of Pdx1-C, for example, is 24% proline, thus a significant fraction of data cannot be collected. While the $^{1}H,^{15}N$-HSQC is typically employed for the initial screening to determine feasibility of project success, for disordered proteins an alternative detection platform should be screened, if proton-based techniques are impractical.

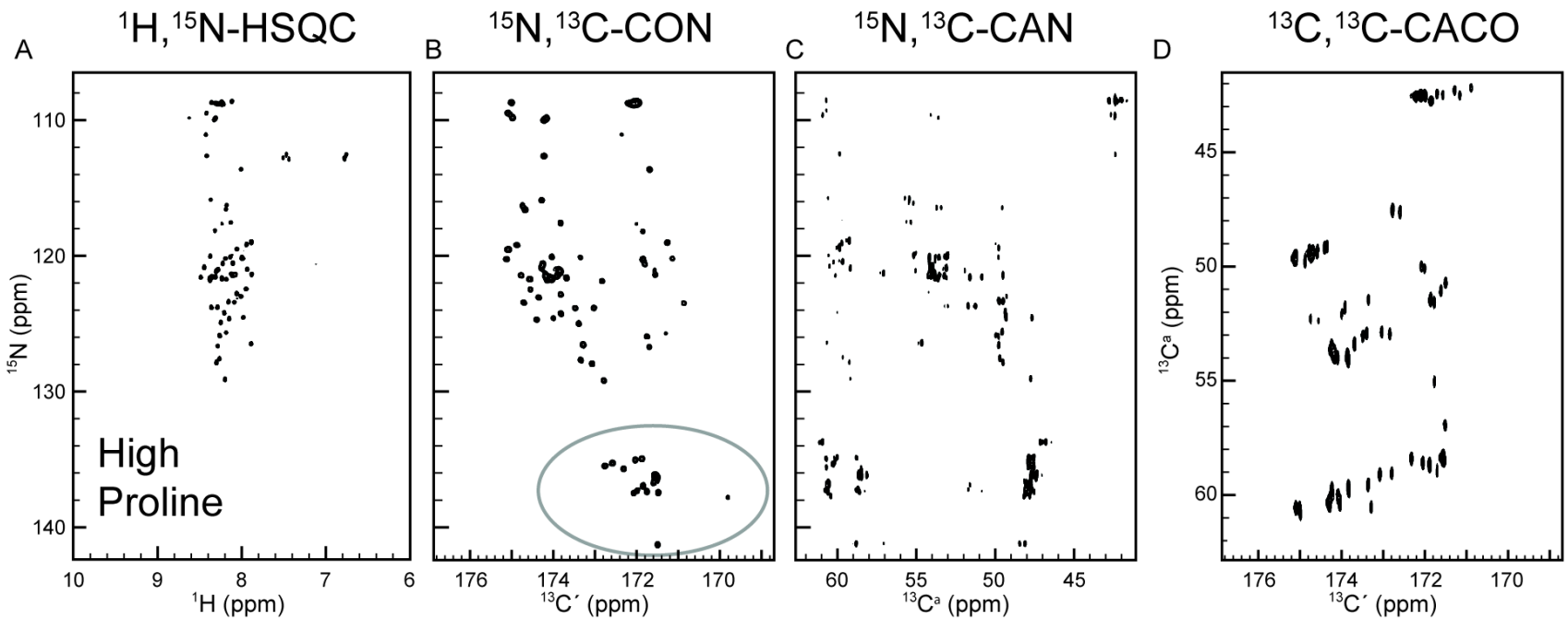

Figure **4-1**. Carbon-detected 2D-NMR experiments are generally effective for generating well resolved spectra of intrinsically disordered proteins. (A-D) Spectra acquired using a 1.3 mM sample of Pdx1-C in 50 mM sodium cacodylate, 50 mM KCl, 5 mM TCEP, pH 6.5, 0.01 % NaN_3, and 10% (v/v) D_2O. The 2D-experiments displayed are (A) ^{1}H,^{15}N-HSQC, (B), ^{15}N,^{13}C-CON, (C) ^{15}N,^{13}C-CAN, and (D) ^{13}C,^{13}C-CACO. All spectra were collected at 298K on a Bruker Avance III spectrometer equipped with a TCI Cryoprobe (^{1}H inner coil), operating at 500 MHz proton resonance frequency (11.7 T).

^{13}C-direct detect methods offer an alternative method to proton based methods, often providing an improvement in signal dispersion for IDPs and detection of proline resonances. These improvements can be seen in Figure 4-1, which reports carbon-detected spectra of Pdx1-C (panel A-D). For ^{13}C-direct detect experiments, there are several platforms to choose from, all of which yield resonances for individual amino acids, including proline residues. The ^{13}C,^{13}C-CACO (Figure 4-1D) reports on the Cα-CO correlation of an individual amino acid and alleviates spectral crowding observed in the ^{1}H,^{15}N-HSQC. While the ^{13}C,^{13}C-CACO remedies most of the spectral crowding observed in the ^{1}H,^{15}N-HSQC, the Cα dimension yields poor dispersion. Similarly, the ^{15}N, ^{13}C-CAN affords improved signal dispersion (Figure 4-1C), in addition to providing intra- and inter-residue correlations. Although these correlations may be useful for backbone assignments using this ^{13}C-platform, long refocusing delays are necessary, which lead to signal loss as a result of relaxation, making the ^{15}N, ^{13}C-CAN ill-suited for backbone assignments. This leads to the ^{15}N, ^{13}C-CON, which correlates the C' of residue *i+1* with the N of residues *i*, yielding improved signal dispersion in both dimensions for IDPs (Figure 4-1B). While providing the best peak dispersion for IDPs, the CON platform can also be used to measure spin relaxation[8], side chain chemical shifts[6], and inter-residue correlation experiments for backbone assignment[3].

Within the CON suite there are three basic experiments: the (C-Start)-CON (Figure 4-2A), (HA-Start)-CON (Figure 4-2B), and the (HN-Start)-CON (Figure 4-2C). Representative spectra collected using this suite of ^{15}N, ^{13}C-CON experiments are shown in Figure 4-2 for Pdx1-C (panels A-C). All spectra reported in Figure 4-2 were collected with 16 scans of signal averaging and 256 increments in the indirect dimension. As can be seen from the 1D projections in Figure 4-2D, the traditional “protonless” carbon-start experiment suffers from reduced signal intensity for a fixed acquisition time, owing to the selection of a low gyromagnetic ratio ^{13}C-

nucleus for the initial excitation (compare the red carbon-start projection to the black and blue proton-start projections). Still, for applications to proteins with paramagnetic metal centers (i.e., many metalloenzymes) the benefits of purely "protonless" spectroscopy should not be discounted. The sensitivity gain associated with proton-start experiments, including both the HA-start (Figure 4-2B) and the HN-start (Figure 4-2C) is pronounced, making these spectra highly attractive for routine applications. For systems possessing a minimal number of proline residues in their sequence, the HN-start CON experiment can be an attractive option and we have used it as the basis for pulse programs to measure, e.g., ^{15}N-spin relaxation with good success.[8] On the other hand, the presence of resonances for ^{15}N, ^{13}C pairs including the amide nitrogen of proline residues makes the (HACA)-CON the pulse program of choice for most routine applications. Of special note, the tolerance of the HA-start format for basic solution conditions (pH > 7.5, discussed in detail below) makes this pulse program especially versatile.

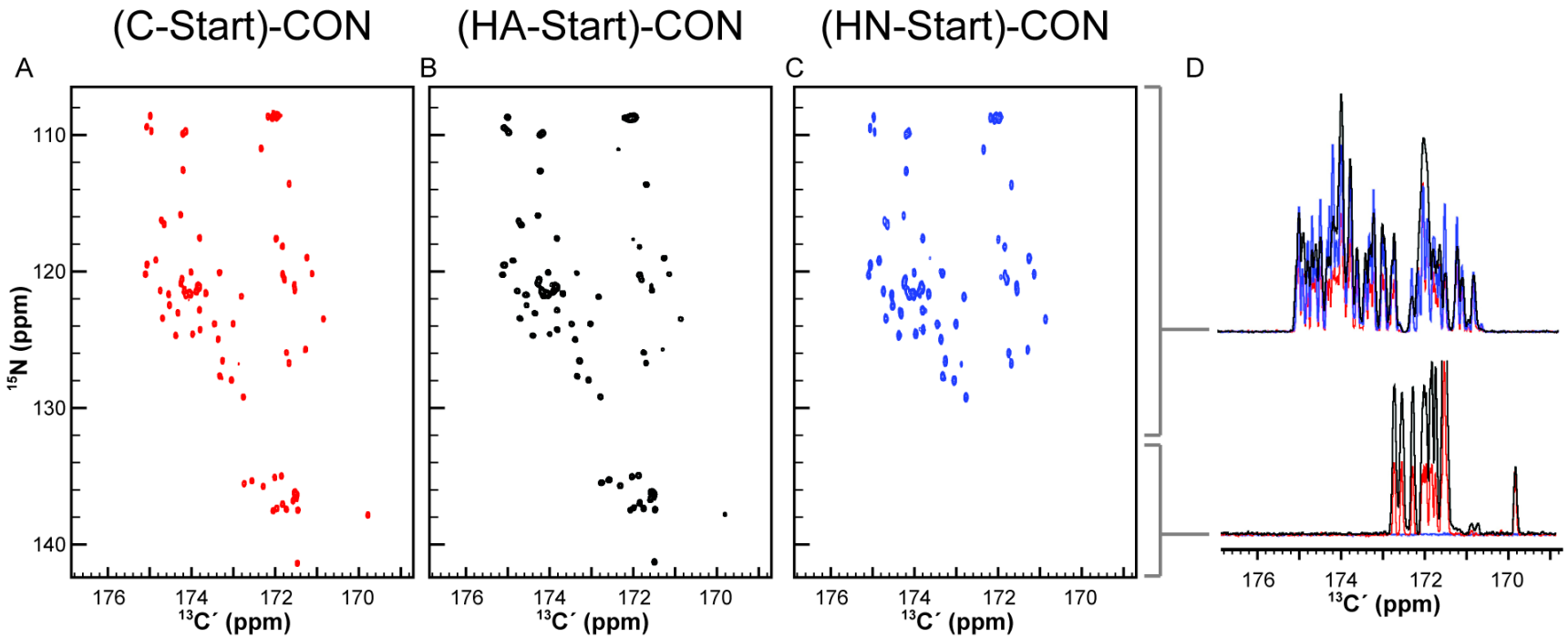

Figure **4-2**. Three common variants of the ^{15}N,^{13}C-CON experiment offer different performance characteristics. Spectra were acquired using (A) C-Start CON, (B) HA-Start CON, and (C) HN-Start CON. Spectra (A-C) were acquired using a 1.3 mM sample of Pdx1-C. Solution conditions were identical to Figure 1. Panels (D) displays the 1D-projection of the non-proline spectral region (105 – 132 ppm; top) and the proline spectral region (132 – 142 ppm; bottom). Both the 2D spectra and their 1D projections are colored to display the C-start data in red, the HA-Start data in black, and the HN-flip data in blue. All spectra were collected at 298K on a Bruker Avance III spectrometer equipped with a TCI Cryoprobe (^{1}H inner coil), operating at 500 MHz proton resonance frequency (11.7 T).

4.2.2 Optimizing magnetic field strength

For classical proton-detected biomolecular NMR experiments, increasing magnetic field strength is often a prudent means to improve spectral resolution. In contrast, most carbon-detected NMR experiments include obligate polarization transfers that require multiple passages through lengthy delays, such as the $^{1}J_{NC}$ refocusing delay. The sensitivity of carbon-detect spectra at higher magnetic field strength, say at 20.0 T compared to the 11.7 T spectrum, is not significantly enhanced, although the linewidths in the direct dimension are modestly improved at higher magnetic field strengths. In our hands, 3D spectra for chemical shift assignment generally yield insufficient signal-to-noise for use when collected at ultra-high field (data not shown). As a general recommendation, moderate magnetic field strengths in the 11.7 – 16.4 T range are more likely to produce reliable spectra for carbon-detected NMR experiments, particularly where ultra-sensitive carbon inner-coil probes are not available.

4.3 Constraints on Sample Conditions

A variety of constraints on data quality must be considered when planning the optimal set of NMR experiments to perform and solution conditions to employ. With protein samples, solubility is often a limiting factor for NMR applications, as concentrations in the 0.1 – 1.0 mM regime are typically needed. All other factors being equal, carbon-detected experiments are generally less sensitive than their proton-detected counterparts, placing an additional burden on the investigator to reach high sample concentrations when carbon-detection is to be applied. Experienced bio-NMR spectroscopists are also familiar with practical constraints on solution conditions, including total salt restrictions with cryogenically cooled probes and pH restrictions

when detection is achieved through exchangeable protons. In this section, we will discuss practical guidelines for sample conditions that are likely to yield high-quality carbon-detected bio-NMR spectra. In general, we find that the restrictions on buffer composition and pH are far less stringent than for proton-detected spectroscopy, although the lower limit for sample concentration is more restrictive, absent carbon inner-coil probes.

4.3.1 Protein concentration requirements

Traditional proton detected experiments yield higher resolution spectra owing to the high sensitivity attained by the large gyromagnetic ratio of the proton nuclei. Protein concentrations for 2 dimensional proton experiments tend to be on the order of 50-200 μM for high quality spectra with the use of cryogenic probes. On the contrary, carbon detected experiments require much higher concentrations to attain a good quality spectra because of the lower gyromagnetic ratio of carbon. Proton detected experiments are much more sensitive than carbon detect experiments, therefore, the concentration of sample should not be undervalued. Based on experience, sample concentrations should be > 500 μM for 2D carbon-detect experiments and > 1 mM for triple resonance experiments. As a result of long delays and extensive magnetization transfers, carbon-detect triple resonance experiments are demanding of sample concentration.

4.3.2 Buffer composition

The stability of IDPs at high concentrations (≥ 500 μM) is crucial if carbon-detected NMR is to be employed for their study. As is common for all proteins, many IDPs are only stable under certain conditions (e.g., high salt, acidic or basic pH, presence of glycerol) that are not necessarily compatible with high-resolution NMR measurements, making buffer composition an

important control point for experimental design. Our research program has been built around several IDPs that require widely varying buffer compositions to reach the concentrations needed. Here we will survey our findings, which are best summarized by reassuring the interested spectroscopist that concentration requirements are, by far, the most restrictive factor in designing an IDP-based research program around carbon-detected NMR.

Many bio-NMR spectroscopists will be familiar with the long held caution that cryogenic probes create severe restrictions on total salt concentration allowable, because tuning and matching becomes challenging as salt concentration grows above >> 500 mM. As many IDPs are highly charged polyampholytes or polyelectrolytes, some require higher than normal salt concentrations to reach stability. In our experience, few if any negative effects are seen on carbon-detected experiments at total monovalent salt concentrations in excess of 1.0 M. If salt concentration does become a concern, then reverting to traditional "protonless" carbon-detected strategies presents a viable work-around.

Salt content is not the only buffer consideration that matters when designing bio-NMR solution conditions. Solvent-exposed backbone amides are subject to efficient proton-exchange with solvent and the rate of exchange is enhanced in mildly basic conditions (pH 7.5 – 8.5), compared to the rate of exchange under mildly acidic conditions (pH 5.5 – 6.5). As IDPs tend to have more solvent-exposed backbones than their cooperatively folded counterparts, the number of amides subject to efficient chemical exchange is dramatically increased, leading to well documented difficulties applying bio-NMR spectroscopy to IDPs that require basic solution conditions to reach sufficient solubility. The loss of resonances to chemical exchange with solvent is readily observed in Figure 4-3, where $^{1}H,^{15}N$-HSQC spectra, collected in buffer at pH 8.5 are presented for Pdx1-C (Figure 4-3A). Comparison with the equivalent spectra in Figure 4-1A, collected at pH 6.5, respectively, the loss of nearly all resonances from the spectra is readily evident. In contrast, $^{15}N,^{13}C$-(HACA)CON spectra of Pdx1-C (Figure 4-3B) collected at pH 8.5

are of almost identical spectral quality to those collected at neutral or mildly acidic pH. Therefore, so long as ^{13}C-start or aliphatic ^{1}H-start experiments are selected, carbon-detected NMR strategies should be robust to nearly any conditions of pH needed to ensure the solubility and stability of the NMR construct.

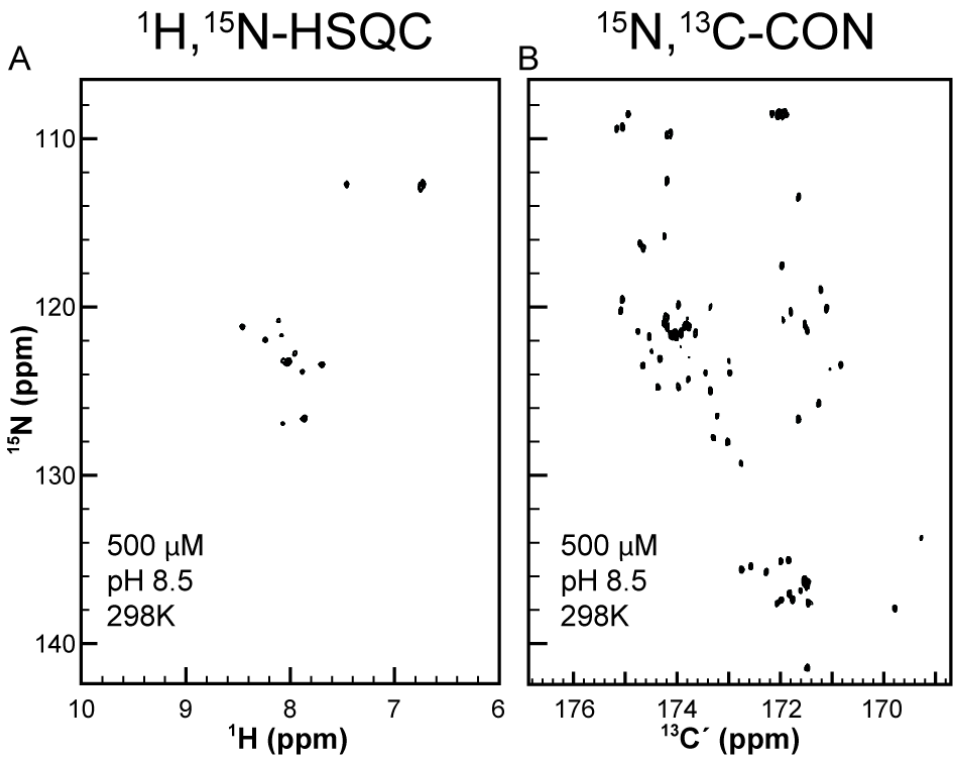

Figure **4-3**. Intrinsically disordered proteins that are only soluble under basic pH conditions are suitable for investigation by carbon-detected NMR. (A) ^{1}H,^{15}N-HSQC spectra and (B) ^{15}N,^{13}C-(HACA)CON spectra. (A, B) Pdx1-C spectra collected on a 500 µM sample in 50 mM Tris, 50 mM KCl, 5 mM TCEP, pH 8.5. All spectra were collected at 298K on a Bruker Avance III spectrometer equipped with a TCI Cryoprobe (^{1}H inner coil), operating at 500 MHz proton resonance frequency (11.7 T).

Finally, so long as they are not isotope-enriched, additives such as glycerol or biological buffers including Tris, HEPES, and other popular choices should also impose no limitations on signal quality. Carbon-detected NMR is resilient to high salt concentrations and a wide range of

pH, although $^{1}H^{N}$-start experiments are subject to many of the same limitations as $^{1}H,^{15}N$-HSQC spectroscopy. In summary, the buffer composition needed to achieve a highly concentrated sample for carbon-detected bio-NMR of disordered proteins should rarely be a limiting factor.

4.4 Chemical Shift Assignment Strategies

Unambiguous chemical shift assignment is crucial for interpretation of all NMR spectra, but this step is not always trivial for IDPs. Accordingly, multi-dimensional carbon-detect approaches have been developed for backbone resonance assignment of IDPs, which are often more generally successful than traditional proton-detected strategies for these systems. In this section, we discuss the experimental suites available, the effects of basic solution pH on some of the available pulse programs, and finally our approach to assigning the backbone resonances of IDPs under basic solution conditions. While learning a new experimental paradigm may seem daunting at first, the reality is that carbon-detected NMR protocols for, e.g., chemical shift assignment, employ many of the same strategies as their proton-detected counterparts. In most cases, pairs of experiments designed to build nearest-neighbor pairs of spin systems are collected for the purpose of "walking" along the backbone.

4.4.1 High dimensional approaches

Recently, non-uniform sampling has become a popular strategy to reduce the experimental time needed for otherwise lengthy high-resolution multidimensional experiments. These time savings have made it possible to measure 4D and 5D carbon-detected experiments for the backbone assignment of IDPs, which have gained increasing attention in the recent literature.[9,

[10] These approaches offer an appealing alternative to the strategy we review here and the interested reader is encouraged to seek more information in the references cited above.

4.4.2 An efficient method for proteins requiring basic pH

Proton-detection methods suffer from restrictions brought about by experimental conditions, such as high salt effects and basic pH (> 7.5) problems, often necessary for the solubility of some IDPs. Carbon-direct detect methods offer a convenient strategy to circumvent experimental condition constraints, while exploiting the enhanced sensitivity from enhancement polarization transfer from protons. For example, we have previously reported a strategy that combines $^{1}H^{N}$-flip N(CA)CON and N(CA)NCO 3D-spectra with a suite of carbon-detected amino-acid specific 2D spectra in an efficient protocol for chemical shift assignment of IDPs solubilized under mildly acidic conditions.[3] Moreover, we have modified the previous pulse sequences to (HACA)-start 3D N(CA)CON and N(CA)NCO, which incorporate polarization transfer from aliphatic protons. Representative strip-plots for Pdx1-C are provided in Figure 4-4. In the (HACA)N(CA)CON spectrum, the unidirectional *i* to *i-1* of the amide nitrogen and carbonyl carbon of each residue and the amide nitrogen of the preceding amino acid are correlated. To alleviate ambiguity and each chemical shift assignment of IDPs, the bi-directional (HACA)N(CA)NCO experiment correlates the amide nitrogen and carbonyl carbon of each residue to the amide nitrogen of the *i-1* and *i*+1 residues. Using a 30% sparse non-uniform sampling schedule on the 3D (HACA)N(CA)NCO, both of these 3D experiments can be collected in a total of ~5 spectrometer-days, allowing cost-effective acquisition.

Of particular note for proline-rich IDPs, when backbone chemical shift assignments are generated from the (HACA)-start 3D experiments, unambiguous assignment through proline residues becomes straightforward. Even for Pdx1-C, for which 24% of the amino acid residues in

our construct are proline, use of the strategy we have discussed here readily yielded nearly complete backbone chemical shift assignment.[3] In combination with carbon-detected amino-acid specific 2D spectroscopy (so-called CAS-NMR[11]) these two spectra offer a straightforward strategy for backbone chemical shift assignment of IDPs under nearly any set of reasonable solution conditions. Further acquisition of 3D C_CCCON[5] and C_H(CC)CON[6] spectra yields complete aliphatic chemical shift assignment as well, providing a robust starting point for structure, dynamics, and functional studies of the target IDP.

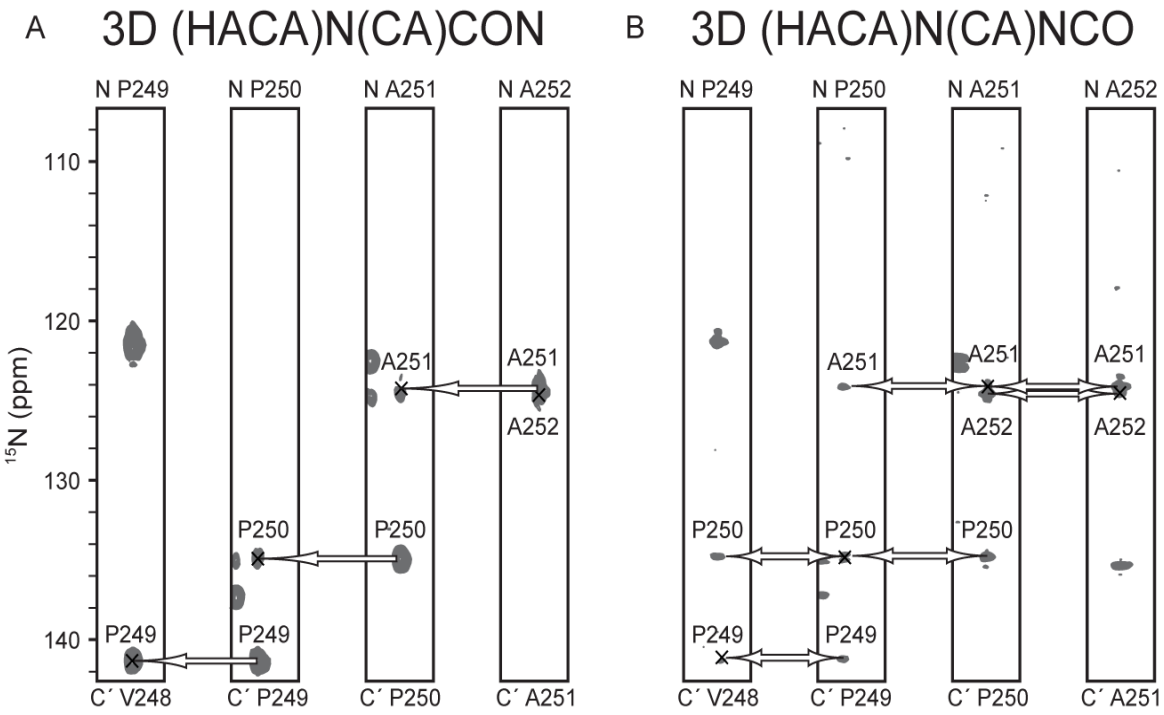

Figure **4-4**. Carbon-detected NMR provides a simple method for backbone assignments of proline-rich intrinsically disordered proteins and proteins soluble under basic pH conditions. (A) (HACA)N(CA)CON strip plots and (B) (HACA)N(CA)NCO strip plots corresponding to 1 mM Pdx1-C. Solution conditions matched those in Figure 3. The (HACA)N(CA)CON spectra were collected with 64 x 128 increments in the indirect dimension and 16 scans of signal averaging. The (HACA)N(CA)NCO spectra were collected using 30% non-uniform sampling with 64 x 128 increments in the indirect dimensions and 32 scans of signal averaging. All spectra were collected at 298K on a Bruker Avance III spectrometer equipped with a TCI Cryoprobe (^{1}H inner coil), operating at 500 MHz proton resonance frequency (11.7 T).

4.5 Conclusions

NMR is a powerful tool for probing the structure and function of IDPs. The carbon-detect experiments discussed here and in the cited work provide an efficient method for generating IDP chemical shift assignments as a pre-requisite for structure-function studies. In this review, we have discussed best practices for establishing viable experimental conditions and selecting pulse programs to overcome pH-related losses of spectral quality. This protocol should be both efficient and highly robust to the conditions interested investigators are likely to encounter. Our recommendation is that spectroscopists begin their carbon-detected studies of IDPs by optimizing the 2D ^{15}N,^{13}C-CON for their system, where the IDP must be soluble and stable at a minimum concentration of 500 μM; in general, it is less important to worry about pH or total salt concentration. Once these conditions are met, established pulse programs or the new (HACA)-start assignment strategy introduced here may be used to rapidly begin what is sure to be an exciting new project.

4.6 Acknowledgements

This work was supported by an NSF-CAREER award (MCB-0953918) to SAS and an NIH predoctoral fellowship (F31GM101936) to MB. All spectra were collected in the Lloyd Jackman NMR Facility in the Department of Chemistry at the Pennsylvania State University. We thank Dr. Emmanuel Hatzakis for assistance with the instruments

4.7 References

1. Jensen, M. R., Ruigrok, R. W., and Blackledge, M. (2013) Describing intrinsically disordered proteins at atomic resolution by NMR, *Curr. Opin. Struct. Biol. 23*, 426-435.

2. Showalter, S. A. (2014) Intrinsically Disordered Proteins: Methods for Structure and Dynamics Studies, *EMagRes 3*, 181-189.

3. Sahu, D., Bastidas, M., and Showalter, S. A. (2014) Generating NMR Chemical Shift Assignments of Intrinsically Disordered Proteins using Carbon-Detected NMR Methods, *Anal. Biochem. 449*, 17-25.

4. Bermel, W., Felli, I. C., Kümmerle, R., and Pierattelli, R. (2008) 13C Direct-detection biomolecular NMR, *Concepts in Magnetic Resonance Part A 32A*, 183-200.

5. Bermel, W., Bertini, I., Felli, I. C., Piccioli, M., and Pierattelli, R. (2006) ^{13}C-detected protonless NMR spectroscopy of proteins in solution, *Prog. Nucl. Magn. Reson. Spectrosc. 48*, 25-45.

6. O'Hare, B., Benesi, A. J., and Showalter, S. A. (2009) Incorporating ^{1}H-chemical shift determination into ^{13}C-direct detected spectroscopy of intrinsically disordered proteins in solution, *J. Magn. Reson. 200*, 354-358.

7. Bermel, W., Bertini, I., Felli, I. C., and Pierattelli, R. (2009) Speeding Up 13C Direct Detection Biomolecular NMR Spectroscopy, *J. Am. Chem. Soc. 131*, 15339-15345.

8. Lawrence, C. W., and Showalter, S. A. (2012) Carbon-Detected N-15 NMR Spin Relaxation of an Intrinsically Disordered Protein: FCP1 Dynamics Unbound and in Complex with RAP74, *J. Phys. Chem. Lett. 3*, 1409-1413.

9. Novacek, J., Haba, N. Y., Chill, J. H., Zidek, L., and Sklenar, V. (2012) 4D non-uniformly sampled HCBCACON and (1)J(NCalpha)-selective HCBCANCO experiments for the sequential assignment and chemical shift analysis of intrinsically disordered proteins, *J. Biomol. NMR 53*, 139-148.

10. Bermel, W., Bertini, I., Felli, I. C., Gonnelli, L., Kozminski, W., Piai, A., Pierattelli, R., and Stanek, J. (2012) Speeding up sequence specific assignment of IDPs, *J. Biomol. NMR 53*, 293-301.

11. Bermel, W., Bertini, I., Chill, J., Felli, I. C., Haba, N., Kumar, M. V. V., and Pierattelli, R. (2012) Exclusively heteronuclear (13) C-detected amino-acid-selective NMR experiments for the study of intrinsically disordered proteins (IDPs), *Chembiochem 13*, 2425-2432.

Chapter 5

Structural and Functional Studies of the Intrinsically Disordered C-terminal Domain of Pdx1

[Preliminary NMR data contributing to this chapter were published in "Generating NMR chemical shift assignments of the intrinsically disordered proteins using carbon-detected NMR methods" by Debasish Sahu, Monique Bastidas, and Scott A. Showalter in *Anal. Biochem.*, (2014) 449, 17-25.]

Abstract

Homeodomain proteins typically have regions of disorder at one or both termini. For example, Pdx1, a homedomain transcription factor involved in maintenance of the pancreas, has long regions of disorder (> 80 residues) at both termini, which have been studied biochemically, but quantitation of their structure and function has been overlooked. To gain a better understanding of the structure and function of Pdx1, we studied the individual C-terminus of Pdx1 using biophysical methods, such as NMR, CD and ITC. These studies reveal that Pdx1-C adopts a random coil structure in solution. Furthermore, we quantify the binding of Pdx1-C and SPOP-Math, and present a model identifying the region involved in protein-protein interactions.

5.1 Introduction

The central dogma of structural biology describes the structure-function relationship of a protein, in which its structure leads to a particular function. As such, this has led to decades of effort aimed at elucidating protein structures from which function of a protein or domain can be

deduced. This view has typically been illustrated by classical examples of well-folded proteins or domains. While this still holds true, a new paradigm has been incorporated into the dogma of structural biology. In 1999, Wright and Dyson presented this new paradigm, in which the structure of a protein is not only described by secondary and tertiary structure, but also random coil or intrinsically disordered.[1] Proteins lacking secondary and tertiary structure are commonly referred to as intrinsically disordered proteins (IDPs).

Motivated by a newfound interest in the study of the configuration of IDPs, structural biologists and bioinformaticists uncovered characteristics in the amino acid sequence of IDPs that correlate with the manifestation of disorder.[2-4] The amino acid sequence of IDPs is enriched in charged and polar residues, while depleted in hydrophobic residues. Many IDPs also have sequences that are rich in one or few residues, such as the poly-glutamine repeat of the androgen receptor N-terminal domain or the proline-rich domain of Sos1.[5, 6] In general, these proteins lack core-forming hydrophobic residues, losing the ability to form stable tertiary structures and, often, secondary structures. Over the past two decades, a plethora of studies and information has been acquired for IDPs, providing substantial evidence of their roles in biological processes or diseased states.[5, 7, 8] IDPs often undergo a disorder-to-order transition upon binding to a partner, have preformed structures that are populated in the unbound state, or are described by a statistical coil model.[9]

While IDPs have been identified in many biological pathways, including signaling, and replication, IDPs have most prominently been identified in the process of transcription.[10] Comparing a large dataset of the amino acid sequence of transcription factors to well-folded proteins, their amino acid frequencies follow the trend observed for IDPs. This strongly suggests that transcription factors have significant regions of disorder. Liu et *al*. show that the number of IDPs with regions of disorder greater than 50 residues is much higher in eukaryotic transcription

factors than prokaryotic transcription factors.[10] This follows the trend, in which disorder increases from prokaryotes to eukaryotes.

One class of transcription factors, homeodomain proteins, is predicted to display a significant degree of disorder. Homeodomain proteins contain a well-folded α-helical motif that binds to DNA sequence-specifically. However, homeodomain folds are often found proximal to regions of disorder at the N-terminus, C-terminus, or both. In particular, the pancreatic duodenal homeobox 1 (Pdx1) homeodomain transcription factor has long regions (> 80 residues) of disorder flanking the termini of the homeodomain. In this study, we focus on the structural characteristics and function of the C-terminus of Pdx1 (Pdx1-C).

Previous and preliminary studies of the structure of Pdx1 point to a conformation represented by a statistical coil model.[11] We expand on and corroborate the structure of Pdx1-C using nuclear magnetic resonance (NMR) methods and circular dichroism (CD). While the previous study of Pdx1-C discussed the chemical shift analysis, which reports on the local environment of the protein, we complete the set of chemical shifts to include chemical shift information for the long stretch of prolines and glycines. Collectively, the structural studies point to a highly dynamic and disordered protein in solution. The role of the coil structure adopted by Pdx1-C has yet to be elucidated, as is the molecular mechanism for its function within a physiological context.

Previously, biochemical studies have pointed to the involvement of Pdx1-C in protein-protein interactions (PPI) with the speckle-type POZ protein (SPOP), which is a well known substrate-adaptor protein for ubiquitin ligase complexes.[12-14] SPOP is composed of three domains - a Math domain, a BTB domain, and a nuclear localization signal domain. Forming part of a larger ubiquitin ligase complex, the SPOP-BTB domain dimerizes to bind to the E3 ubiquitin ligase while the SPOP-Math domain recruits protein substrates for ubiquitination by the E3/E2 ubiquitin-protein ligase complex. Using *in vitro* interaction assays, a moderate interaction

between Pdx1-C and the SPOP-Math domain was identified, while a weaker interaction with the BTB domain was observed.[13] The exact residues of the C-terminal domain of Pdx1 were not identified, nor were the binding interactions quantified, nevertheless the binding region of the Math domain was narrowed down. The residues of Pdx1-C involved in binding to the SPOP-Math domain are in a conserved region (aa 210-238).[13] Interaction of ubiquitin ligase, SPOP, and Pdx1 is mediated by the Math domain and Pdx1-C, thereby regulating Pdx1 levels and subsequently the expression of key β-cell hormones known to be activated by Pdx1. Furthermore, ubiquitination of Pdx1 was enhanced in the presence of low glucose levels (~ 5mM), and repressed under high glucose concentrations (~ 25 mM), when increased levels of insulin are necessary.[15] Interestingly, a mutation within the conserved Pdx1-C sequence that interacts with SPOP results in the loss of a stable interaction between the Math domain and the C-terminus of Pdx1. This underscores the importance in atomistically defining the interaction and understanding how this E224K mutation impacts the β-cell phenotype. Thus, Pdx1-C plays a regulatory role in insulin transcription by its involvement in protein-protein interactions critical for regulation of transcription.[13, 16] Here we provide direct biochemical evidence for the interaction between Pdx1-C and the SPOP-Math domain. By calorimetric methods, the interaction strength of Pdx1-C with the SPOP-Math domain is quantified. The results are consistent with previously reported binding constants of the Math domain with other validated substrates.[14] Furthermore, the binding region of Pdx1-C is reported using NMR. Finally, we provide insight for the function of this long disordered region of Pdx1.

5.2 Materials and Methods

5.2.1 Preparation and Purification of Pdx1-C

A synthetic Pdx1 gene with codons optimized for *Escherichia Coli* was purchased from Geneart and the Pdx1 C-Terminus (amino acids 204-283 of the human sequence; subsequently referred to as Pdx1-C), was subcloned by PCR into pET49b (Novagen) encoding a Glutathione-S-Transferase tag, a 6x His tag and a 3C protease recognition site upstream of the cloning site. The recombinant plasmid was transformed into BL21(DE3) competent cells for protein over-expression. Cell growth conditions and the protein purification protocol for Pdx1-C were identical to our previously reported procedures for the Pdx1-homeodomain,[17] except as noted below. As a final step to ensure full purification, Pdx1-C was subjected to size exclusion chromatography using a HiPrep 26/60 Sephacryl S-200 HR column (GE Life Sciences) in 50 mM Tris pH 7.5, 150 mM NaCl, 5 mM β-mercaptoethanol, and 1 mM EDTA. Following concentration using an Amicon Ultra centrifugal filter device (Millipore) that contained a PES 3000 MWCO membrane, Pdx1-C was buffer exchanged into 50 mM cacodylate, final pH 6.5, 50 mM KCl, and 5 mM TCEP. Protein concentration was determined by Direct Detect FT-IR (Millipore) using the molecular weight of 8089 g/mol.

5.2.2 Preparation and Purification of SPOP-Math

The Math domain of SPOP (aa 1-166 or 28-166) was subcloned from a synthetic human SPOP gene purchased from GeneArt into the pET49b vector using the *XmaI* and *EcoRI* cutsites encoding Glutathione-S-transferase tag and His_6-tag at the N-terminus of the construct. Optimal

growth conditions for maximum expression of SPOP-Math is an overnight induction (OD of 0.8-1.0) at 24 °C for 16-18 hours. The SPOP domain was purified using nickel-affinity, ensuring all buffers contained 5 mM of fresh reducing agent. First, following lysis and centrifugation of 1L growth, the supernatant was passed over nickel resin and the flow through was collected for replication. The resin was washed with 30-40 mL of wash buffer, followed by 10 mL of lysis buffer. The SPOP construct was eluted in 20 mL of elution buffer. The affinity procedure was completed in duplicate or triplicate to maximize protein purification. The elutions were pooled and the tags were cleaved with 3C Protease during an overnight dialysis in 1.8 L followed by a second nickel affinity purification. The tagless protein was passed through the same nickel column for further purification from the tag and protease. The flow through, containing the protein, was passed through a second time after elution and equilibration of the column. The protein was then concentrated using a spin column and buffer exchanged into buffer. UV-Vis spectroscopy was used to determine the concentration by the Beer-Lambert law and the molar absorbance at 278 nm of 29,400 $M^{-1}cm^{-1}$ (reduced disulfides).

5.2.3 Nuclear Magnetic Resonance Spectroscopy

The Pdx1-C concentrations for NMR spectroscopy were approximately 1.0 mM. Backbone chemical shift assignments were generated using samples in 50 mM cacodylate, 50 mM KCl, 5 mM TCEP, adjusted to a final pH of 6.5, 10% D_2O (w/v), and 0.01% NaN_3. To complete backbone assignments of Pdx1-C, carbon-detect NMR methods were used to assign most resonances.[11] All NMR spectra were recorded at 298K on a BruckerAvance III 500 MHz and 600 MHz spectrometers equipped with TCI cryoprobes for increased sensitivity. All NMR spectra were processed by NMRPipe and analyzed with SPARKY (SPARKY 3.113; T. D.

Goddard and D. G. Kneller, University of California, San Francisco, CA). Residual dipolar coupling and relaxation data were analyzed in MATLAB (MathWorks). Backbone chemical shifts for Pdx1-C were deposited in the BMRB (accession no. 19596).

5.2.4 Residual Dipolar Coupling

An unaligned sample of Pdx1-C was prepared for a Shigemi tube to a concentration of 1.1 mM. The same sample was aligned in a neutral 2% polyacrylamide gel medium. The sample window was maintained as close as possible by adjusting the height of the Shigemi tube to 1.6 mm. The neutral 2% polyacrylamide gels were made in 5mm tubes, removed and cut to 1.6 cm length before allowing drying. NMR experiments were processed in NMRPipe and visualized in Sparky. RDC analysis was completed using in-house MATLAB scripts.

5.2.5 ^{15}N Spin Relaxation

Spin relaxation experiments measuring ^{15}N T_1 and T_2 spin relaxation were collected using the ^{13}C-detected pulse programs, CON(T1)-IPAP and CON(T2)-IPAP, which were performed at 298K on a Bruker Avance III 600 MHz spectrometer. T_1 relaxation delays were generated by acquiring a set of spectra with delays, 50, 150, 250, 350, 450, 550, 650, and 750 ms. Similarly, decays for T_2 relaxation were generated by acquiring 8 spectra with delays 16, 48, 80, 112, 144, 176, 208, and 240 ms.

5.2.6 Circular Dichroism Spectroscopy

Further purification of Pdx1-C was necessary for circular dichroism experiments. Following the purification procedure outlined above, the sample was concentrated to a volume ≤ 1.5 mL and further purified by size exclusion using an S100 gel filtration column on an AKTA FPLC system. Because Pdx1-C has an extinction coefficient of zero at A_{280}, fractions 25-35 were run on a gel to determine the exact fractions for spin concentrating. Pdx1-C was buffer exchanged into 50 mM cacodylate, 5 mM TCEP, pH 6.5 and concentrated to 85.4 μM, which was determined by A_{205} measurements. The extinction coefficient used to determine the concentration was 237950 cm^{-1} M^{-1}, which was computed using the method by Anthis and Clore.[18]

Experiments were collected on a Jasco J-1500 instrument at 298K. Using a 0.1 mm cuvette, the buffer measurement was collected followed by subsequent collection of a triplicate set of sample measurements. Experimental parameters included setting the data pitch to 0.025 nm, a scanning speed of 200 nm/min, and a bandwidth of 1 mm. A baseline correction was applied automatically to the experiment. No further quantitative analysis was performed on the data.

5.2.7 Isothermal Titration Calorimetry (ITC)

Titrations were collected in triplicate on a VP-ITC instrument at 298 K. Protein concentrations of ligand (Pdx1-C) ranged from 300-500 μM, while the target macromolecule (SPOP-Math) was present at 12-20 μM, respectively. Samples were co-dialyzed in 100 mM Potassium Phosphate, 100 mM Potassium Chloride, 2 mM TCEP, pH 6.0 overnight at 4 °C. Heats

of dilution were accounted for by subtracting a reference titration of ligand into buffer. The raw data was analyzed using a single site binding model as implemented in ORIGIN 7.0.

5.3 Results and Discussion

5.3.1 Chemical Shift Analysis of Pdx1-C

Following backbone assignments of Pdx1-C, chemical shift analyses were carried out using two programs, Secondary Structure Prediction (SSP) and δ2D.[19, 20] Each program predicts secondary structure populations based on the chemical shifts of nuclei, with respect to random coil or secondary structure libraries, respectively.

Secondary structural elements, including helical content, β-structure, and coil structure are predicted by the SSP software. Positive values signify α-helical structure propensity and β-structure is demonstrated by negative values. Typically, large magnitudes of positive and negative values indicate stable secondary structure, whereas small values, either positive or negative, are indicative of random coil structure. For Pdx1-C, the values obtained are relatively small in magnitude throughout the entire protein sequence, which highlights pronounced coil structure for Pdx1-C (Figure 5-1A).

The sequence of Pdx1-C is very rich in proline and glycine content, ~22% and 15%, respectively. These residues have the propensity to form polyproline II (PPII) structure, thus we hypothesized that Pdx1-C exhibited PPII structure. Without the need to carry out additional experiments, we employed the δ2D program, which also predicts PPII propensities based on chemical shifts, in addition to helix, β-structure, and coil. The results, in Figure 5-1B, from δ2D indicate Pdx1-C adopts a random coil conformation, however, a little β-structure and PPII

structure may be present. Regions exhibiting β-structure includes a polyglycine tract (aa 216-219) and a polyproline tract (aa 241-245), whereas shorter regions are predicted to have PPII structure (aa 221-223 and aa 269-270). Interestingly, part of the region in the predicted binding region by SPOP-Math is predicted to exhibit β-structure.

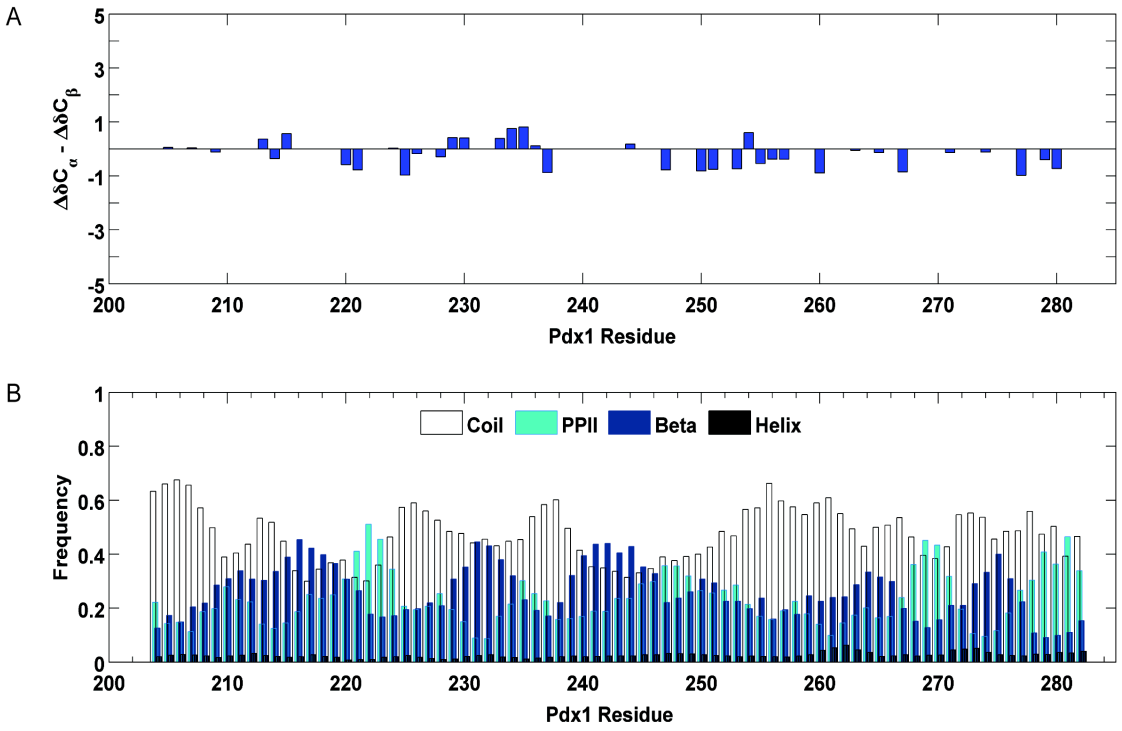

Figure **5-1.** Chemical shift analysis of Pdx1-C using (A) SSP and (B) δ2D programs. In panel A, secondary chemical shifts of Pdx1-C are computed against a referencing library of random coil and secondary structural chemical shifts. δ2D output, in panel B, details the secondary structure propensity on a per residue basis, showing coil propensity (white), polyproline II or PPII (green), β-structure (blue), and α-helix (black).

5.3.2 Structural Insight of Pdx1-C

Residual dipolar coupling (RDC) NMR experiments are one of the several NMR methods to determine secondary structure properties of a protein system.[21-26] A great deal of structural information can be obtained from collecting several RDC experiments. Local structure, as well as

long-range order in proteins can be elucidated by RDC measurements of varying intermolecular vectors. For proteins that exhibit folded secondary structure, such as an α-helix or extended β-sheet structure, RDCs yield characteristic features indicative of folded secondary structure. However, for proteins that do not exhibit folded secondary structures, RDC data tends to be featureless; nevertheless the data provides information about the structure, or lack of folded structure.

Folded regions of proteins exhibiting stable helical or β-sheet structures have pronounced RDC values in regions that encompass these structural elements. Slightly larger amplitudes for the N-H backbone vector are observed even for IDPs, however, RDCs for the N-H vector is generally 2-5 times greater in magnitude for helical or β-sheet structures. Because RDC measurements are robust and sensitive to report on secondary structural elements of proteins, these experiments were carried out for Pdx1-C to identify secondary structural elements, if any existed. RDC of the N-H backbone vector was measured for Pdx1-C in a neutral polyacrylamide alignment medium and are illustrated in Figure 5-2. No structural elements are present in the solution structure of Pdx1-C as evidenced by the lack of large amplitude features indicative of secondary structural elements. Even though RDC measurements are ideal for determining secondary structure of proteins, an additional method for secondary structure comparison and validation was employed.

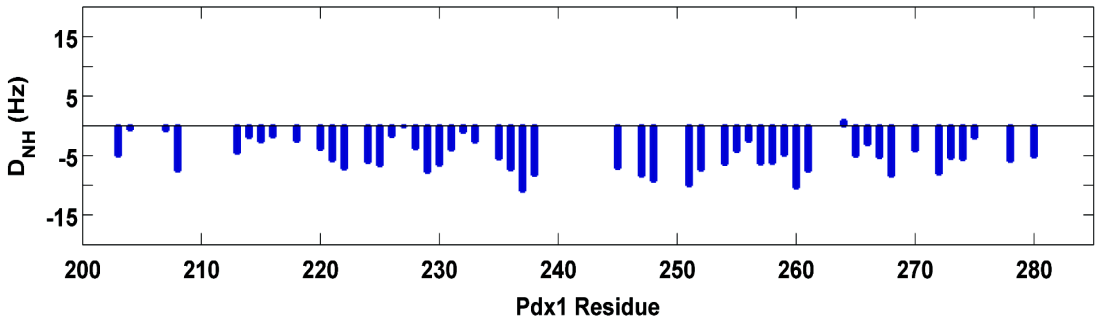

Figure **5-2** N-H residual dipolar coupling (RDC) NMR experiment of Pdx1-C.

Circular dichroism is an excellent technique to probe the secondary structural elements in a protein, such as α-helix, β-sheet, coil, and PPII,. The CD profile of these secondary elements is quite unique. A protein whose structure is composed of various secondary structures will produce a CD spectrum that is a linear combination of the respective secondary structure signatures. Based on secondary structure prediction algorithms and NMR methods presented, Pdx1-C exhibits random coil structure. The far-UV CD spectrum of Pdx1-C, shown in Figure 5-3, illustrates a protein with random coil with a signature minimum near 198 nm.[27, 28] Thus, based on several techniques to probe the structure of Pdx1-C, the protein is primarily statistical coil in solution. However, one cannot rule out the presence of additional secondary structure, such as PPII or β-strand because the local minimum in Figure 5-3 is shifted.

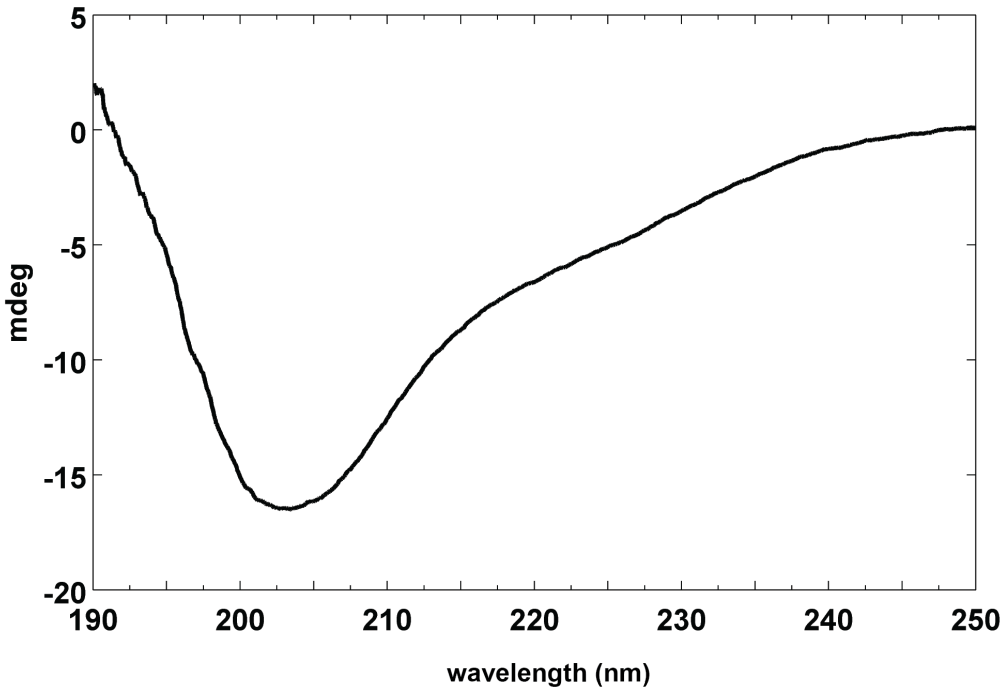

Figure **5-3.** Raw data of the far-UV CD spectrum of Pdx1-C collected in triplicate at 298K.

5.3.3 Dynamics of Pdx1-C

In a cellular environment, globular proteins are highly dynamic objects; the dynamics of disordered proteins are more extreme. Measurement of ^{15}N spin relaxation is a well-established technique used to monitor the conformational dynamics of proteins. Backbone dynamics report on the inherent flexibility of a given amino acid, where greater flexibility yields increased T_1 dynamics. Spin dynamics may illuminate the impact of IDP conformational dynamics to the mechanism of protein-protein interactions mediated by Pdx1. ^{15}N spin relaxation experiments were collected to monitor the backbone dynamics of Pdx1-C and shed light on secondary structural characteristics that may be relevant in protein-protein interactions.

The backbone of Pdx1-C is quite dynamic, as evidenced by the high amplitude global T_1 baseline of approximately 0.7 seconds at 600 MHz. Of particular interest, the poly-glycine region within residues 210-220 and residues 266-273 (a serine rich region) display greater flexibility in both T_1 in comparison to the overall global baseline (Figure 5-4). Typical of highly flexible proteins, such as disordered proteins, the T_1 relaxation properties of residues experiencing large fluctuations are high, compared to 0.4 seconds for T_1 for well-folded domains.[11]

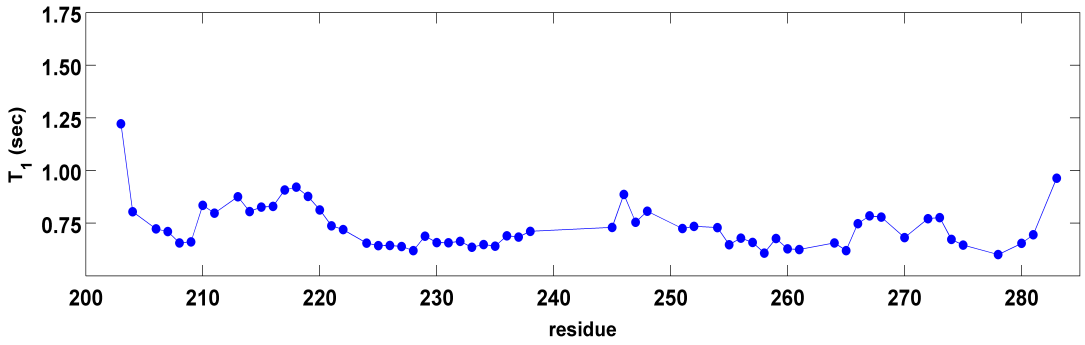

Figure **5-4.** ^{15}N T_1 spin relaxation illustrating the backbone dynamics of Pdx1-C collected at 600 MHz.

In elucidating the structure of Pdx1, we found no evidence to distinguish the conformational state of the C-terminus from a random coil polymer in solution. In terms of the separate question of secondary structure bias, we similarly identify no observable stable or transient secondary structure (Figure 5-4). The view on structure-function of proteins has typically been that well-formed secondary and tertiary structure imparts protein function. Contrary to the classical structure-function relationship, the structure of the C-terminus of Pdx1 is described as a random coil, yet previous reports have demonstrated a functional role in protein-protein interaction responsible for Pdx1 regulation.

5.3.4 Quantitative Analysis of Pdx1-C binding to SPOP-Math

An interaction between SPOP-Math and Pdx1 was determined qualitatively by *in vitro* interaction assays, which showed a moderate interaction between the Math domain of SPOP and Pdx1-C. The consensus binding sequence for SPOP-Math was identified as ϕ-π-S-S/T-S/T, where ϕ is a hydrophobic residue and π is a polar residue. Interestingly, the C-terminus of Pdx1 does not contain the consensus binding sequence reported by Schulman et *al*.[14] Therefore, to determine the integrity of the molecular interaction and identify residues involved interaction NMR was performed for Pdx1-C in the absence and presence of SPOP-Math.

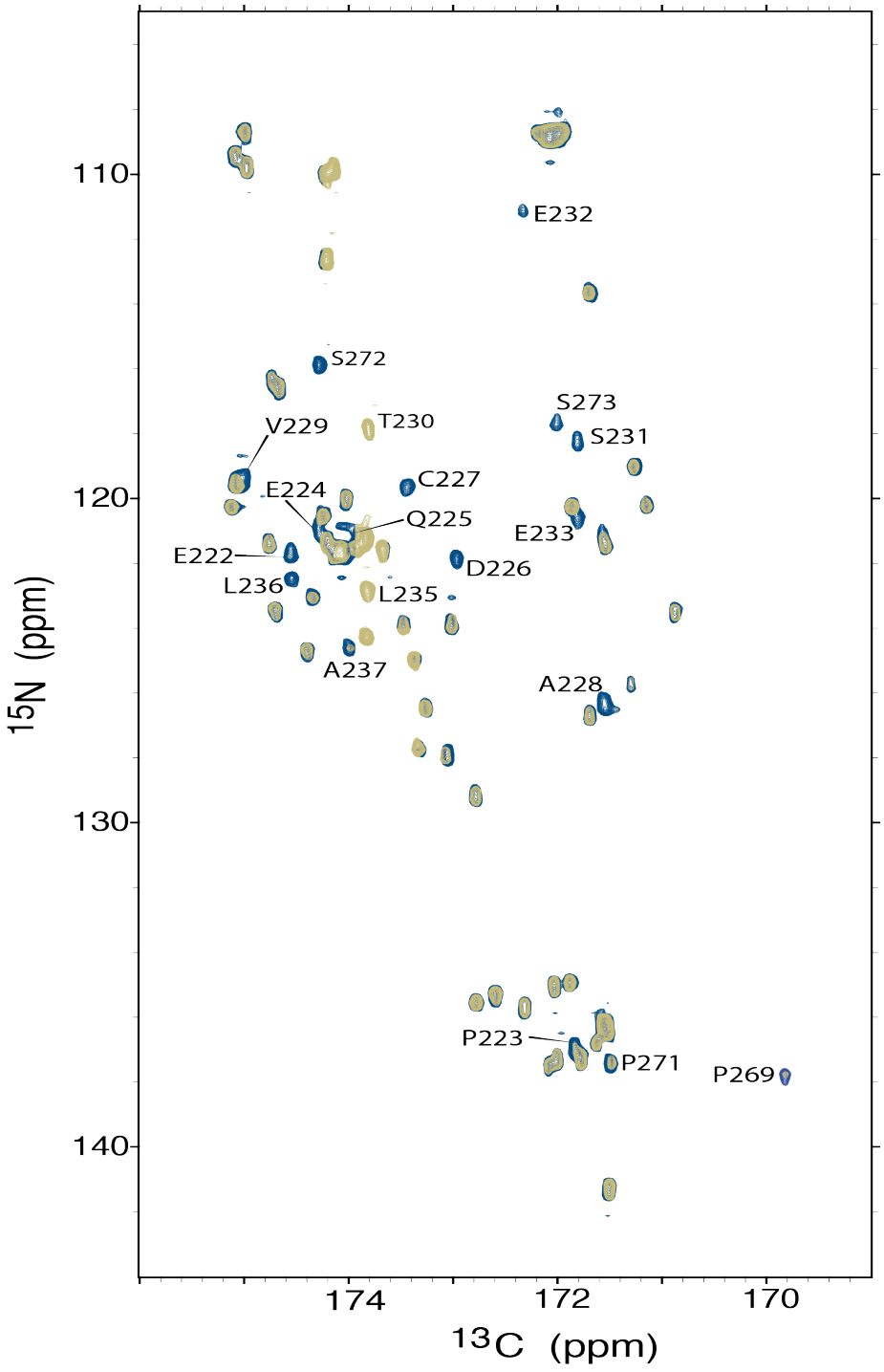

Figure **5-5.** NMR spectra of Pdx1-C in the absence (blue) and presence of SPOP-Math (yellow). Labeled residues indicate residues experiencing line broadening due to binding and are within the region identified essential for SPOP binding.

Because NMR enables detection of individual amino acids and their chemical shift is dependent on the environment, residues involved in SPOP binding will experience a change in their chemical environment resulting in a chemical shift perturbation. When binding is in intermediate exchange, meaning binding occurs within the direct timescale of chemical shifts, the result is the disappearance of peaks upon binding. Therefore, with NMR the exact binding epitope of SPOP-Math$^{1-166}$ can be mapped. Figure 5-5 shows the NMR spectra of Pdx1-C in the absence of SPOP-Math$^{1-166}$ (blue) and in its presence (yellow). The disappearance of peaks demonstrates a definitive binding interaction due to intermediate exchange between states or increase in molecular tumbling time due to increase in molecular weight, which leads to severe line broadening or signal loss. Chemical shifts involved in the interaction are mapped to residues 222-236. Liu et *al.* implicated residues 210-238 to be essential for the SPOP-Math domain to inhibit Pdx1 activity.[13] Our NMR data reports on the residues involved in binding and is consistent with the reported binding region. Although we identified more than five residues interact with SPOP, this is likely the result of a dynamic interaction encompassing additional residues. Surprisingly, a secondary binding site encompassing residues 269-273 is observed from the SPOP-bound Pdx1-C NMR spectrum. The finding of the secondary binding site by NMR is consistent with the 2009 study by Zhuang et *al.*, which provides evidence for SPOP (Math-BTB construct) binding to two substrate binding sites.[14] Overall, the NMR study presented is consistent with two SPOP-binding sites within the Pdx1-C sequence from which a structural binding model based on the SPOP-Math co-crystal structure is introduced (Figure 5-6). Further studies to determine the binding face and configuration will be necessary for attaining atomistic details of this interaction.

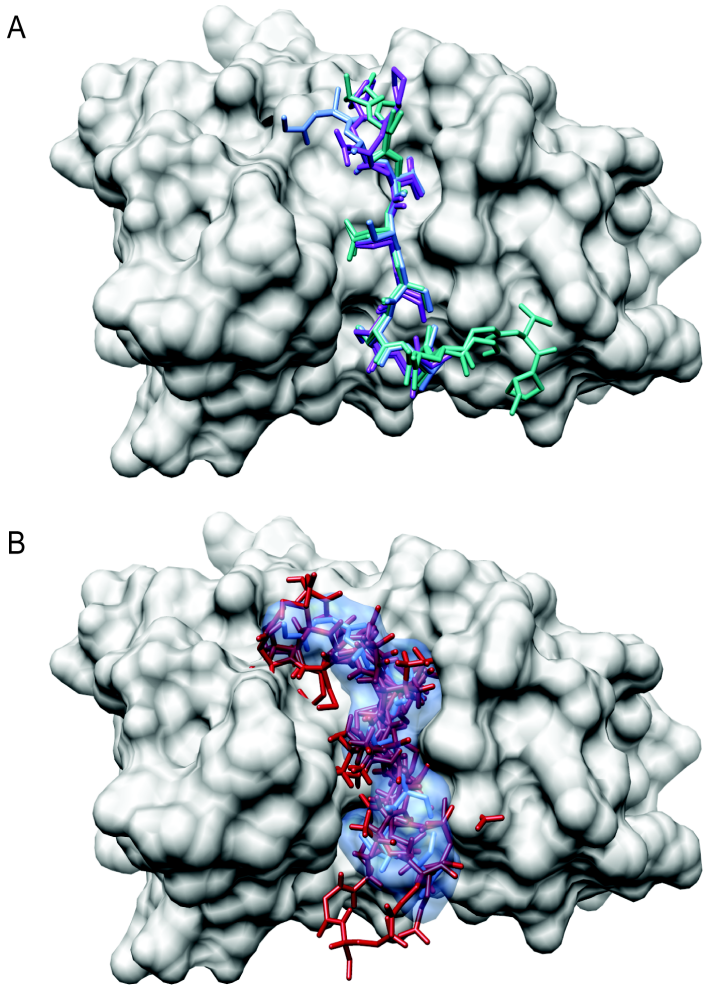

Figure **5-6.** Structure of SPOP-Math:Substrate binding interface. Panel A shows the co-crystal structures of SPOP-Math with various substrates encompassing the canonical binding sequence. In panel B, the deviated Pdx1-C sequence (aa 226-233) modeled into the SPOP-Math binding pocket (red).

After demonstrating an interaction between Pdx1-C and SPOP-Math, a quantitative evaluation of the interaction was carried out by ITC. Many of the structural and functional studies with SPOP-Math were done in the context of the cooperatively folded region of the Math domain encompassing residues 28-166. However, a 27 amino acid disordered N-terminal arm was never incorporated into the studies. Therefore, we chose to work with this length of protein to study the function in the context of the N-terminal arm to, later, elucidate the effect of the arm in substrate

binding, if any. We measured the thermodynamics of Pdx1-C with SPOP-Math$^{1-166}$ (aa 1-166), encompassing the disordered 27 residue N-terminal arm.

Regardless of the divergent consensus binding sequence in Pdx1-C, weak binding to Pdx1-C by SPOP-Math$^{1-166}$ was observed by ITC (Figure 5-7). The interaction produced a thermogram resulting in a binding interaction in the low μM range. The affinity of the interaction was 25 ± 1 μM with a binding enthalpy of 5.8 ± 0.3 kcal mol^{-1} and a stoichiometry of 1.2 ± 0.1. Although the Pdx1 sequence does not contain the exact SPOP consensus binding sequence, the binding constant is within the range determined for SPOP-Math. Binding studies with SPOP-Math$^{28-165}$ and its peptide substrates demonstrated a wide range of affinities. Although the substrates encompass the consensus binding sequence, the range of affinities for the SPOP-Math with different substrates is between 4 μM and 100 μM.[14]

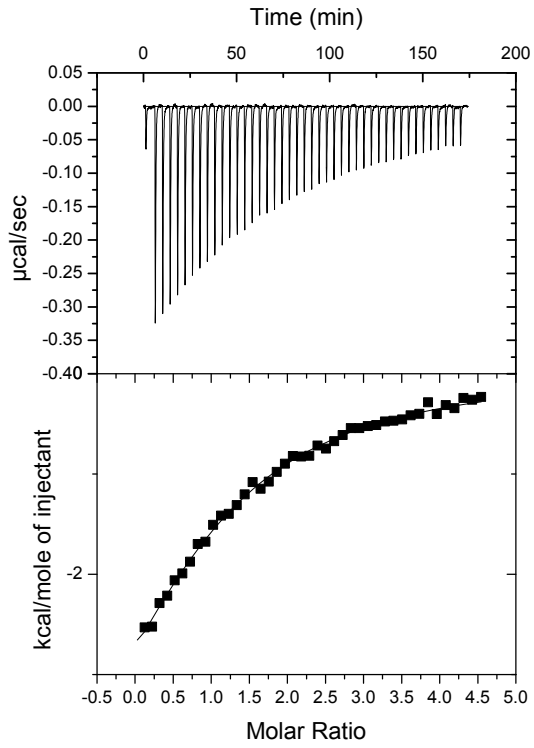

Figure **5-7**. Representative ITC titration of Pdx1-C into SPOP-Math (aa 1-166) at 298K.

Several co-crystal structures of SPOP-Math domain bound to different substrates were solved by X-ray crystallography. A SPOP-Math consensus binding sequence was identified, in which the sequence preference for SPOP-Math is Φ-π-S-S/T-S/T, where Φ is a non-polar residue and π is a polar residue.[14] In the co-crystal structures, hydrophobic and hydrogen bonds stabilize the interaction between the Math domain and its substrate. Do electrostatics play a role in binding, as a result of the positively charged residues in the binding site? The interaction between SPOP-Math and Pdx1-C was first reported to be within residues 210-238. By NMR studies presented in this chapter, I show that the binding region encompasses many of the residues reported earlier by Liu et *al.*[13] However, the amino acid sequence of Pdx1-C within this region does not match exactly the SPOP-Math consensus sequence reported by Zhuang et *al.*[14] Perhaps, the Pdx1-C binding region presents a new SPOP binding sequence. A co-crystal structure of SPOP-Math with a Pdx1-C peptide and a Pdx1-C E224K peptide would be an exciting avenue to pursue, especially because of the novelty of the Pdx1-C sequence for SPOP binding. This structure would define the interactions between the two proteins at an atomistic level.

Furthermore, a MODY mutation is present in the binding region of Pdx1-C, E224K, which likely presents charge repulsions with SPOP-Math, likely precluding binding. Liu et *al.* show that this Pdx1-C mutation prevents SPOP from inhibiting Pdx1 transactivation ability. Binding studies of the full-length Pdx1-C mutant and peptides with SPOP-Math will provide insight into the mechanism of their interaction, which may be purely electrostatic as a result of the mutation, and how it this interaction, or lack thereof leads to a negative cellular response.

5.4 Conclusion

Extensive structural and dynamics studies of apo Pdx1-C reveal a completely disordered protein in solution and this coil structure is deemed important for protein-protein interaction with the SPOP protein. The SPOP-Math domain binds to a consensus amino acid sequence ϕ-π-S-S/T-S/T, where ϕ is a hydrophobic residue and π is a polar residue, however, the binding epitope within Pdx1-C is divergent. We quantified the interaction between SPOP-Math and Pdx-C, which yield a binding affinity in the micromolar region. Finally, using NMR, the epitope was mapped to residues 225-233. In summary, our studies reveal the random coil structure of Pdx1-C in the apo state, enabling an interaction between SPOP-Math.

5.5 Acknowledgements

I would like to acknowledge Dr. Scott A. Showalter for providing Figure 5-6 and Lai Shi for his diligence and inquiry while I trained him on isothermal titration calorimetry and for preparing and setting up the Pdx1-C/SPOP-Math titrations. Dr. Debashish Sahu wrote the in-house MATLAB scripts used to analyze the RDC data.

5.6 References

1. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *Journal of molecular biology 293*, 321-331.

2. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *Journal of molecular graphics & modelling 19*, 26-59.

3. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins-Structure Function and Genetics 42*, 38-48.

4. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins 41*, 415-427.

5. Davies, P., Watt, K., Kelly, S. M., Clark, C., Price, N. C., and McEwan, I. J. (2008) Consequences of poly-glutamine repeat length for the conformation and folding of the androgen receptor amino-terminal domain, *Journal of molecular endocrinology 41*, 301-314.

6. McDonald, C. B., Bhat, V., Kurouski, D., Mikles, D. C., Deegan, B. J., Seldeen, K. L., Lednev, I. K., and Farooq, A. (2013) Structural landscape of the proline-rich domain of Sos1 nucleotide exchange factor, *Biophysical chemistry 175-176*, 54-62.

7. Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011) Intrinsically disordered proteins: regulation and disease, *Current opinion in structural biology 21*, 432-440.

8. Uversky, V. N., Dave, V., Iakoucheva, L. M., Malaney, P., Metallo, S. J., Pathak, R. R., and Joerger, A. C. (2014) Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases, *Chemical reviews 114*, 6844-6879.

9. Gibbs, E. B., and Showalter, S. A. (2015) Quantitative biophysical characterization of intrinsically disordered proteins, *Biochemistry 54*, 1314-1326.

10. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors, *Biochemistry 45*, 6873-6888.

11. Sahu, D., Bastidas, M., and Showalter, S. A. (2014) Generating NMR chemical shift assignments of intrinsically disordered proteins using carbon-detected NMR methods, *Analytical biochemistry 449*, 17-25.

12. Liu, A., Desai, B. M., and Stoffers, D. A. (2004) Identification of PCIF1, a POZ domain protein that inhibits PDX-1 (MODY4) transcriptional activity, *Molecular and cellular biology 24*, 4372-4383.

13. Liu, A., Oliver-Krasinski, J., and Stoffers, D. A. (2006) Two conserved domains in PCIF1 mediate interaction with pancreatic transcription factor PDX-1, *FEBS letters 580*, 6701-6706.

14. Zhuang, M., Calabrese, M. F., Liu, J., Waddell, M. B., Nourse, A., Hammel, M., Miller, D. J., Walden, H., Duda, D. M., Seyedin, S. N., Hoggard, T., Harper, J. W., White, K. P., and Schulman, B. A. (2009) Structures of SPOP-substrate complexes: insights into molecular architectures of BTB-Cul3 ubiquitin ligases, *Molecular cell 36*, 39-50.

15. Claiborn, K. C., Sachdeva, M. M., Cannon, C. E., Groff, D. N., Singer, J. D., and Stoffers, D. A. (2010) Pcif1 modulates Pdx1 protein stability and pancreatic beta cell function and survival in mice, *The Journal of clinical investigation 120*, 3713-3721.

16. Mosley, A. L., and Ozcan, S. (2004) The pancreatic duodenal homeobox-1 protein (Pdx-1) interacts with histone deacetylases Hdac-1 and Hdac-2 on low levels of glucose, *The Journal of biological chemistry 279*, 54241-54247.

17. Bastidas, M., and Showalter, S. A. (2013) Thermodynamic and structural determinants of differential Pdx1 binding to elements from the insulin and IAPP promoters, *Journal of molecular biology 425*, 3360-3377.

18. Anthis, N. J., and Clore, G. M. (2013) Sequence-specific determination of protein and peptide concentrations by absorbance at 205 nm, *Protein science : a publication of the Protein Society 22*, 851-858.

19. Camilloni, C., De Simone, A., Vranken, W. F., and Vendruscolo, M. (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts, *Biochemistry 51*, 2224-2231.

20. Marsh, J. A., Singh, V. K., Jia, Z., and Forman-Kay, J. D. (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation, *Protein science : a publication of the Protein Society 15*, 2795-2804.

21. Bouvignies, G., Markwick, P. R., and Blackledge, M. (2007) Simultaneous definition of high resolution protein structure and backbone conformational dynamics using NMR residual dipolar couplings, *Chemphyschem : a European journal of chemical physics and physical chemistry 8*, 1901-1909.

22. De Biasio, A., Ibanez de Opakua, A., Cordeiro, T. N., Villate, M., Merino, N., Sibille, N., Lelli, M., Diercks, T., Bernado, P., and Blanco, F. J. (2014) p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins, *Biophysical journal 106*, 865-874.

23. Mohana-Borges, R., Goto, N. K., Kroon, G. J., Dyson, H. J., and Wright, P. E. (2004) Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings, *Journal of molecular biology 340*, 1131-1142.

24. Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M., and Bax, A. (1997) Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution, *Nature structural biology 4*, 732-738.

25. Vajpai, N., Strauss, A., Fendrich, G., Cowan-Jacob, S. W., Manley, P. W., Grzesiek, S., and Jahnke, W. (2008) Solution conformations and dynamics of ABL kinase-inhibitor complexes determined by NMR substantiate the different binding modes of imatinib/nilotinib and dasatinib, *The Journal of biological chemistry 283*, 18292-18302.

26. Valafar, H., Mayer, K. L., Bougault, C. M., LeBlond, P. D., Jenney, F. E., Jr., Brereton, P. S., Adams, M. W., and Prestegard, J. H. (2004) Backbone solution structures of proteins using residual dipolar couplings: application to a novel structural genomics target, *Journal of structural and functional genomics 5*, 241-254.

27. Kjaergaard, M., Norholm, A. B., Hendus-Altenburger, R., Pedersen, S. F., Poulsen, F. M., and Kragelund, B. B. (2010) Temperature-dependent structural changes in

intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?, *Protein science : a publication of the Protein Society 19*, 1555-1564.

28. Rath, A., Davidson, A. R., and Deber, C. M. (2005) The structure of "unstructured" regions in peptides and proteins: role of the polyproline II helix in protein folding and recognition, *Biopolymers 80*, 179-185.

Chapter 6

Preliminary Studies Leading to Future Directions

Abstract

Pdx1 is an essential transcription factor required for differentiation of the pancreas and duodenum, and maintaining the β-cell phenotype. Pdx1 activates transcription of several genes of β-cell hormones by binding sequence specifically to dsDNA via its homeodomain motif. However, the majority of the primary sequence of Pdx1 is disordered. Few functional studies of these regions of Pdx1 have been conducted; however, much knowledge is to be gained by structurally, thermodynamically, and functionally studying these regions. Thus, this chapter is divided into several sections based on Pdx1 construct, 1) Pdx1-N, 2) Pdx1-HDC, and 3) *in vivo* studies with Pdx1 and human promoter elements. Specifically, each section contains an introduction, preliminary data, as well as a discussion and future directions.

6. General Introduction

Many biological processes require the association of a complex or several complexes to carry out specific activities, for example, transcriptional activation of a gene, which requires the association of the RNA Polymerase II complex. For transcriptional activation of *insulin*, Pdx1 binds to specific DNA sites (referred to as A-boxes), while other transcription factors bind to sites adjacent to Pdx1 target sites. These proteins interact synergistically to activate transcription, by participating in DNA binding and recruitment of other proteins and/or the transcription initiation complex. The overall mechanism for transcriptional activation involving Pdx1 is not fully

understood. In the full length Pdx1 protein the homeodomain is embedded within longer disordered regions, 145 and 77 residues N- and C-terminal to its DNA binding motif, respectively. Given the highly dynamic nature of disordered regions, these likely play a role in Pdx1 gene activation by protein-protein interactions. The disordered regions of Pdx1 have been shown biochemically to interact with other proteins, supporting their role in gene regulation.

6.1.1 Introduction – Pdx1-N

The Pdx1 homeodomain is at the core of an insulin activation complex also composed of E47 and Beta2.[1] Beta2 and E47 belong to the basic helix-loop-helix (bHLH) motif class and represent a heterodimer between the broadly expressed E47 and the cell-type specific partner Beta2. bHLH proteins form dimers at the HLH interface to collectively bind E box promoter elements, GC-rich sequences. Figure 6.1-1 illustrates the binding of the Beta2/E47 dimer binding to E2 site, while the Pdx1-HD interacts with the A3 site on the human *insulin* gene promoter. The results of Ohneda, *et. al.*[1] demonstrate that the homeodomain is involved in the protein-protein interactions with Beta2:E47. However, their homeodomain construct length encompasses a stretch of ~11 disordered residues N-terminal to the homeodomain, which may be involved in the protein-protein interactions with the proximal bHLH dimer (Figure 6.1-1).

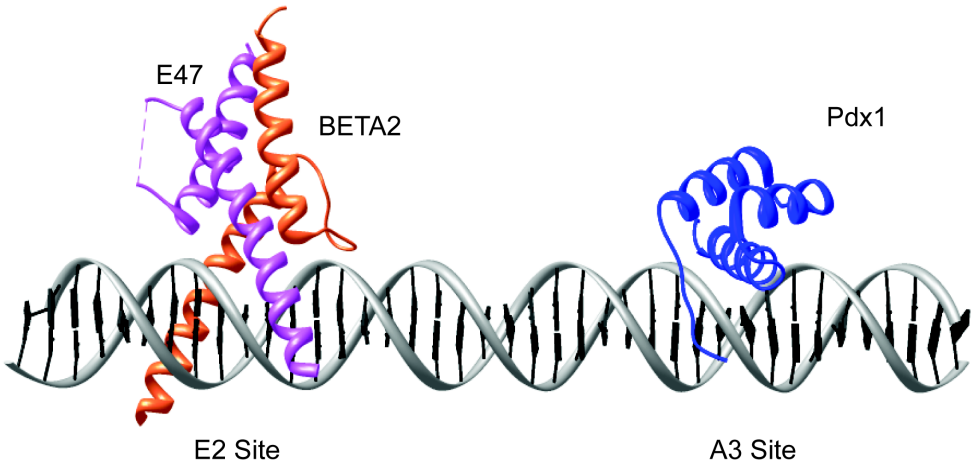

Figure **6.1-1.** Depiction of a hypothetical Insulin promoter showing the E2 and A3 elements bound by the respective activators E47/BETA2 and Pdx1.

In another study, recruitment of and interaction with histone acetyltransferase p300 by Pdx1 and the Beta2:E47 complex has been demonstrated.[2] Specifically, p300 interacts with the transactivation domain of Pdx1, which is mapped to the N-terminal residues 1-79.[3] This in turn leads to cooperative recruitment of the RNA polymerase II pre-initiation complex by protein-protein interactions with the insulin activation complex. Previous reports reveal the formation of Pdx1-protein interactions synergistically activates insulin transcription.[1, 2] However, neither the precise identity of the binding interfaces, structures of the complexes, nor thermodynamic data have ever been determined in a way that enables construction of a rational mechanistic model for the assembly of the insulin activation complex.

6.1.2 Materials and Methods

6.1.2.1 Expression and Purification of Pdx1-N

A synthetic *pdx1* gene was purchased from Geneart and the N-terminus of Pdx1 (aa 1-146), was subcloned by PCR into pET49b (Novagen) encoding a Glutathione-S-transferase (GST) fusion tag, a 6x His tag and a 3C protease recognition site upstream of the cloning site. The recombinant plasmid was transformed into BL21 (DE3) competent cells for protein over-expression. The cells were incubated at 30 °C until an OD of 0.5-0.7 was reached and expression induced with a final IPTG concentration of 0.5 mM. Cells were harvested 4 hours post-induction and lysed by sonication at 4 °C. The suspension was centrifuged at 4 °C for 30 min at 11500 rpm in a Beckman Coulter Allegra 25R using a TA-14-50 rotor. The supernatant was passed over a Ni-NTA (Novagen) column, and the protein was eluted with imidazole (200 mM). The His_6-tag was cleaved using 3C protease at 4°C overnight while also dialyzing away the imidazole. The contents of the dialysis bag were passed over the same nickel column and the protein was eluted

by imidazole gradient. Pdx1-N eluted with an imidazole concentration starting at 10 mM to 50 mM.

6.1.3. Preliminary Results

6.1.3.1 NMR Spectra of Pdx1-N

Based on secondary structure prediction software, the N-terminus of Pdx1 is predicted to be primarily disordered. To test this prediction, we turned to NMR spectroscopy, which can quickly report on the relative structure of a protein. Chemical shifts are natural reporters of their chemical environment. For example, for well-folded proteins, a typical $^{1}H,^{15}N$-HSQC will yield good signal dispersion, however, for IDPs, signal dispersion is weak and chemical shifts center around 8 ppm. Therefore, a 2D correlation of the amide nitrogen and amide proton for Pdx1-N was collected using an $^{1}H,^{15}N$-HSQC pulse sequence . However, in hind sight based on the size of the protein, a more appropriate method would have been a $^{1}H,^{15}N$-TROSY pulse sequence. Figure 6.1-2, nevertheless, illustrates the characteristics of a disordered protein for Pdx1-N. Overall, the protein can be purified to carry out experiments, but the conditions must be optimized for each experiment.

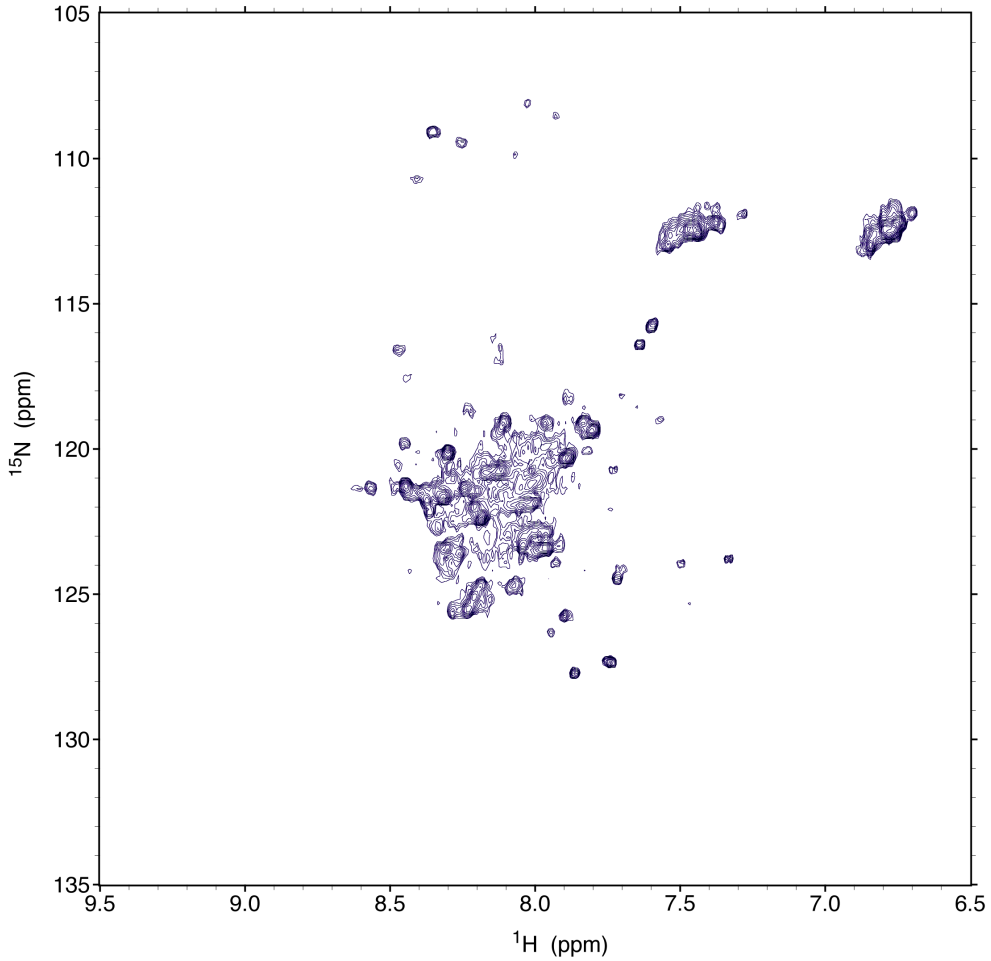

Figure **6.1-2.** 2-dimensional ^{15}N,^{1}H-HSQC spectrum of Pdx1-N (aa 1-146) at 298K.

6.1.4. Discussion and Future Directions

6.1.4.1 Biophysical Characterization and Functional Studies of Pdx1-N

The N-terminal domain of Pdx1 encompasses the transactivation domain (aa 1-79), which is required to regulate gene expression.[3] The N-terminal tail is involved in protein-protein interactions with other transcription factors, which engage in synergistic activation of β-cell genes. These interactions have been characterized by mutational analysis and *in vitro* interaction assays in a qualitative manner. Studying these interactions by thermodynamics will provide an

understanding of the mechanism of binding. Furthermore, the *in vitro* interaction assays only encompass the well-folded domains of these proteins, which are shown to interact. However, these proteins, including Pdx1 at the N-terminus have regions of disorder, which may interact.

In contrast to Pdx1-C, the primary sequence of the N-terminus of Pdx1 contains a number of hydrophobic and aromatic residues, which are less frequent in disordered regions. In addition, Pdx1-N also has a high number of proline residues. This leads me to hypothesize that the N-terminus of Pdx1 exhibits residual structure that may be involved in interactions with other transcription factors. Polyproline helix II structure is commonly found in globular proteins. However, supporting speculation that intrinsically disordered proteins adopt PPII structure, an increasing number of studies show that PPII may be a dominant structure in 'random coil' proteins. For example, the structure of the proline-rich (PR) domain of Sos1 was interrogated by circular dichroism (CD) in the presence of increasing concentrations of urea or guanidium-HCl.[4, 5] The native structure of the PR domain exhibited random coil structure, however, interestingly, the protein demonstrated significant structural changes with increasing denaturant. These structural changes represented a loss in random coil structure and a gain in PPII for the chemically denatured state. Conversely, in another study temperature-dependent CD of three different IDPs was measured. The formation or loss of secondary structure can be measured by temperature-dependent CD studies. This particular study showed that a loss of PPII structure was observed for all three IDPs considered.[6] Not only are intrinsically disordered proteins involved in many biological processes, PPII structure is connected with processes, for example, signal transduction, cell motility, and transcription.[7] To further investigate the structural properties of Pdx1-N, temperature-dependent and denaturant-dependent CD studies would illuminate the structural characteristics and ultimately relate it to its function.

6.1.4.2 Structural and Functional Investigation MODY Mutations

Six of the 12 clinically documented mutations (5 missense and 1 frameshift) of Pdx1 are located within its N-terminus.[8-14] The N-terminus encompasses the transactivation domain (TAD), for which four of the 6 mutations are within this TAD. Do these mutations affect the structure of this region or the protein or do they affect the functionality of this domain? It is likely one or more of the mutations affect structure and binding, which leads to a loss in the ability to participate in protein-protein interactions. Structural studies of the wild-type and mutant forms will elucidate the detrimental effects these mutations have on Pdx1 activity.

6.1.5 References

1. Ohneda, K., Mirmira, R. G., Wang, J. H., Johnson, J. D., and German, M. S. (2000) The homeodomain of PDX-1 mediates multiple protein-protein interactions in the formation of a transcriptional activation complex on the insulin promoter, *Mol Cell Biol 20*, 900-911.

2. Qiu, Y., Guo, M., Huang, S. M., and Stein, R. (2002) Insulin gene transcription is mediated by interactions between the p300 coactivator and PDX-1, BETA2, and E47, *Mol Cell Biol 22*, 412-420.

3. Peshavaria, M., Henderson, E., Sharma, A., Wright, C. V., and Stein, R. (1997) Functional characterization of the transactivation properties of the PDX-1 homeodomain protein, *Molecular and cellular biology 17*, 3987-3996.

4. McDonald, C. B., Bhat, V., Kurouski, D., Mikles, D. C., Deegan, B. J., Seldeen, K. L., Lednev, I. K., and Farooq, A. (2013) Structural landscape of the proline-rich domain of Sos1 nucleotide exchange factor, *Biophysical chemistry 175-176*, 54-62.

5. McDonald, C. B., Seldeen, K. L., Deegan, B. J., and Farooq, A. (2008) Structural basis of the differential binding of the SH3 domains of Grb2 adaptor to the guanine nucleotide exchange factor Sos1, *Archives of biochemistry and biophysics 479*, 52-62.

6. Kjaergaard, M., Norholm, A. B., Hendus-Altenburger, R., Pedersen, S. F., Poulsen, F. M., and Kragelund, B. B. (2010) Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?, *Protein science : a publication of the Protein Society 19*, 1555-1564.

7. Rath, A., Davidson, A. R., and Deber, C. M. (2005) The structure of "unstructured" regions in peptides and proteins: role of the polyproline II helix in protein folding and recognition, *Biopolymers 80*, 179-185.

8. Al-Quobaili, F., and Montenarh, M. (2008) Pancreatic duodenal homeobox factor-1 and diabetes mellitus type 2, *International Journal of Molecular Medicine 2008*, 399.

9. Hani, E. H., Stoffers, D. A., Chevre, J. C., Durand, E., Stanojevic, V., Dina, C., Habener, J. F., and Froguel, P. (1999) Defective mutations in the insulin promoter factor-1 (IPF-1) gene in late-onset type 2 diabetes mellitus, *J Clin Invest 104*, R41-48.

10. Hansen, L., Urioste, S., Petersen, H. V., Jensen, J. N., Eiberg, H., Barbetti, F., Serup, P., Hansen, T., and Pedersen, O. (2000) Missense mutations in the human insulin promoter factor-1 gene and their relation to maturity-onset diabetes of the young and late-onset type 2 diabetes mellitus in caucasians, *J Clin Endocrinol Metab 85*, 1323-1326.

11. Macfarlane, W. M., Frayling, T. M., Ellard, S., Evans, J. C., Allen, L. I., Bulman, M. P., Ayres, S., Shepherd, M., Clark, P., Millward, A., Demaine, A., Wilkin, T., Docherty, K., and Hattersley, A. T. (1999) Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes, *J Clin Invest 104*, R33-39.

12. Stoffers, D. A., Ferrer, J., Clarke, W. L., and Habener, J. F. (1997) Early-onset type-II diabetes mellitus (MODY4) linked to IPF1, *Nat Genet 17*, 138-139.

13. Weng, J., Macfarlane, W. M., Lehto, M., Gu, H. F., Shepherd, L. M., Ivarsson, S. A., Wibell, L., Smith, T., and Groop, L. C. (2001) Functional consequences of mutations in the MODY4 gene (IPF1) and coexistence with MODY3 mutations, *Diabetologia 44*, 249-258.

14. Gragnoli, C., Lindner, T., Chiaromonte, F., Colasurdo, L., Dantonio, T., Gragnoli, F., Gragnoli, G., Manetti, E., Signorini, A. M., and Marozzi, G. (1998) Identification of a missense mutation (P33T) in the insulin promoter factor-1 (IPF-1) gene in an Italian patient with early onset type 2 diabetes, *Diabetologia 41*, 412-412.

6.2.1 Introduction – Pdx1-HDC

Transcription requires numerous proteins involved in DNA-binding, protein-protein interactions, and post-translational modifications to carry out activation of a gene. As such, the structure of transcription factors not only encompasses well-folded domains, but these proteins also have regions of disorder that can modulate binding interactions involving DNA or proteins. Disordered regions can participate in multiple interactions, enabled by their inherent plasticity.[1] Intrinsic disorder in transcription factors of various types is crucial for transactivation by interaction with other protein factors, substrate recognition, or DNA-binding inhibition.[1] Based on bioinformatics studies and our NMR studies (Chapter 5), Pdx1-C adopts an entirely random coil structure. Is this region only involved in Pdx1-protein interactions or does this tail have a role in DNA-binding by the homeodomain. In an effort to study the role of Pdx1-C in homeomdomain binding, I generated a Pdx1-HDC construct and carried out ITC studies with a panel of native promoter elements used in Chapters 2 and 3. The studies demonstrated a slight increase in binding affinity by the homeodomain. Interestingly, the binding resulted in a different binding isotherm capturing an additional event, which may be a protein conformational change, change in the oligomerization state, or perhaps another explanation. Unfortunately, the second event captured by ITC was not a priority at the time and was not pursued for further investigation.

6.2.2. Materials and Methods

6.2.3.1 Preparation and Purification of HDC

A recombinant plasmid was generated as described above for SPOP. The resulting plasmid was transformed into protein expressing BL21 (DE3) cells where the cells were grown and induced at 37 °C for 4 hours with 250 µL of 1M IPTG per 500 mL. The cells were lysed at 4

°C by sonication for 12 cycles of 20 seconds continuous pulse and a 1 minute rest period. Purification by nickel affinity was completed with all buffers containing 5 mM reducing agent as described above for SPOP purification. The molar absorbance used to calculate protein concentration by UV-Vis spectroscopy at 278 nM is 14,000 $M^{-1}cm^{-1}$. Double labeled $^{13}C,^{15}N$ protein was grown and expressed in M9 minimal media under the same conditions as unlabeled protein.

6.2.3.2 Isothermal Titration Calorimetry

Pdx1-HDC and dsDNA were co-dialyzed overnight in 100 mM Cacodylate, pH 7.3, 100 mM KCl, and 2 mM β-mercaptoethanol. Samples were centrifuged at 4200 rpm for 5-10 minutes to pellet precipitation or solid particles, which may clog the instrument. Pdx1-HDC (120 μM) was titrated into dsDNA (10 μM) in a series of titrations at 7.5 μL injection volume, with the exception of the first injection, which was set to 2 μL. Data analysis was performed using the Origin software. The last 5 points of a titration were averaged and simple math was performed to complete baseline subtraction.

6.2.3. Preliminary Results

6.2.3.1 Structural Studies of Pdx1-HDC

In an attempt to do NMR studies on Pdx1-HDC, I collected a $^{15}N,^{1}H$-HSQC spectrum to determine the feasibility of studying this protein using NMR methods. In the spectra shown in Figure 6.2-1, the spectrum is well dispersed for an intrinsically disordered protein tethered to a well-folded domain. Many resonances display sharp line widths, but a decent number of

resonances display broad line shapes, which likely indicate their participation in some exchange process such as dimerization or binding. There is a single cysteine residue in the C-terminus of Pdx1 that is exposed and can easily form a disulfide bond if there is insufficient reducing agent present in the buffer.

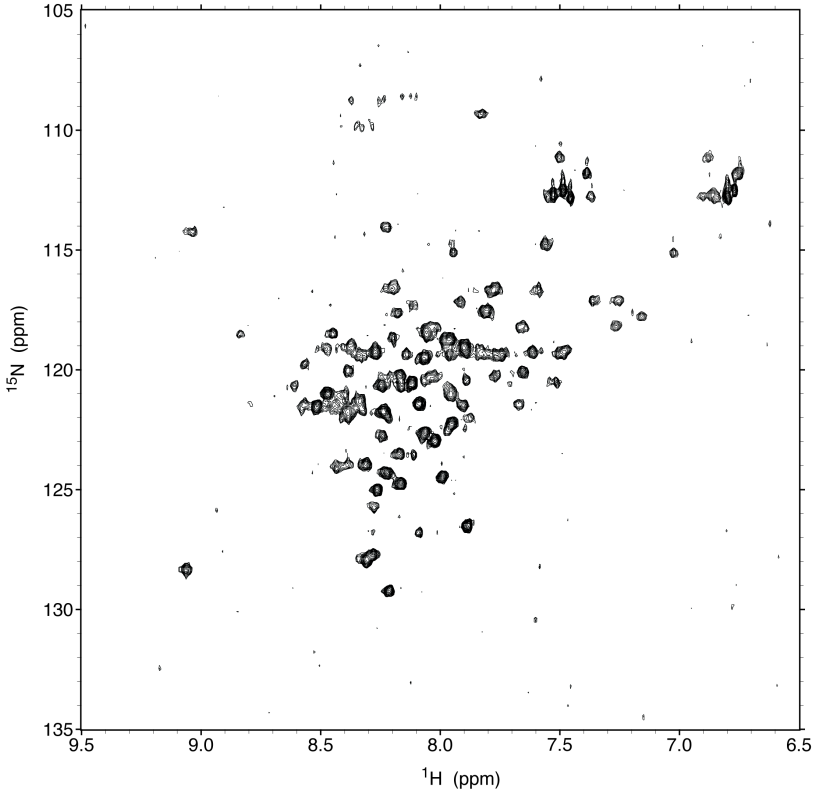

Figure **6.2-1.** ^{15}N,^{1}H-HSQC spectrum of Pdx1-HDC (aa 146-283) at 298K.

6.2.3.2 Functional Studies of Pdx1-HDC

A series of titrations with Pdx1-HDC with a panel of DNA sequences was completed at 25 °C on the VP-ITC. The ITC data is presented in Table 6.2-1, along side the data for Pdx1-HD for comparison. The data presented supports the notion that IDPs act as affinity tuners; here the disordered C-terminus of Pdx1 enhances the binding potential of Pdx1's homeodomain. For all

duplex DNA sequences, Pdx1-HDC bound more tightly perhaps due to the relatively high frequency of positively charged residues that could be making contacts with the DNA.

Table **6.2-3.** Calorimetric data for the Pdx1-HDC (aa 146-283) with a panel of promoter-derived and consensus-derived sequences.

		Homeodomain C-terminus	
Sample Name	**Sequence (5′-3′)**	**K_d (nM)**	**ΔH (kcal/mol)**
Insulin A1	AGG CCC TAA TGG GCC A	5.2 ± 0.3	-10.44 ± 0.02
Insulin A3	AGA CTC TAA TGA CCC G	7.4 ± 0.9	-11.87 ± 0.04
IAPP A1	GGA AAT TAA TGA CAG A	4.4 ± 2.1	-8.58 ± 0.03
Consensus	CCA CTC TAA TGA GTT C	5.6 ± 0.3	-10.54 ± 0.02
Consensus 5′G	CCA CTG TAA TGA GTT C	5.8 ± 0.3	-9.59 ± 0.03
Consensus 5′A	CCA CTA TAA TGA GTT C	6.2 ± 1.0	-7.63 ± 0.03
Consensus 5′T	CCA CTT TAA TGA GTT C	8.9 ± 0.8	-6.52 ± 0.02
Consensus 3′T	CCA CTC TAA TTA GTT C	4.3 ± 1.2	-11.13 ± 0.02
Consensus 3′A	CCA CTC TAA TAA GTT C	21.5 ± 2.1	-5.48 ± 0.03
Consensus 3′C	CCA CTC TAA TCA GTT C	13.4 ± 1.0	-10.36 ± 0.02

6.2.4. Discussion and Future Directions

The long disordered C-terminal tail of Pdx1 may be interacting with the DNA via two positively charged patches (aa 207-209 or aa 277-283) enhancing the affinity of Pdx1. The mechanisms by which this is occurring is unknown and warrants further investigation. Furthermore, the interaction of Pdx1 with a subset of DNA sequences resulted in an additional event being measured by ITC. It is possible that the dissolution of a dimer, followed by binding can result in two events. Further studies of Pdx1-HDC to study aggregation at high concentration required for ITC should be evaluated. In an attempt to separate the DNA-binding event from the

second event, a preliminary temperature-dependent study was carried out to try to separate the two events. Instead, however, a large and negative change in heat capacity was calculated from the binding enthalpies at three temperatures, which is consistent with specificity binding of a protein to DNA. Nevertheless, the two events were unable to be separated. Several exciting avenues can be pursued, in which the temperature-dependent study, the separation of the two events, and determination of the second event can be pursued. Though some new data was acquired by studying this construct, the particular protein was relatively unstable. Prior to beginning additional studies with this construct, buffers conditions stabilizing this construct will need to be tested.

6.2.5 References

1. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors, *Biochemistry 45*, 6873-6888.

6.3.1 Introduction – Cell-Based Studies of Pdx1 Function

Pdx1 is well known to bind A-boxes of several promoter regions of multiple genes, including *insulin*, *islet amyloid polypeptide* (*iapp*), *glucose transporter 2*, and *glucokinase*. To maintain glucose homeostasis, insulin is expressed and secreted under high concentrations of glucose. Whereas, IAPP is a regulatory hormone that inhibits insulin secretion at high concentrations and slows gastric emptying.[1, 2] Studies show that insulin and IAPP hormones are co-secreted, albeit in differing concentrations. Although insulin concentrations are (25-50 times) higher than IAPP in β-cells, their relative concentrations can vary within and across β-cells.[1, 2] Still, the mechanism by which insulin and IAPP levels are held at unequal abundance is not well documented. Pdx1 being a key player in glucose homeostasis remains at the core of understanding the variance in insulin and IAPP levels. In Chapter 2, the data demonstrates a preference of promoter element sequence targeted by Pdx1. Based on calorimetric data presented in Chapter 2, Pdx1 binds preferentially to insulin elements over IAPP elements, specifically with a two-fold increase in *in vitro* binding affinity. Although this is a relatively small difference, the binding *in vivo* may have cooperative effects upon addition of cis acting elements and position within the promoter.[3-6] Does this binding preference and affinity translate to Pdx1 behavior within a cellular environment?

Although the Pdx1 interactions with varying promoter elements have been studied using *in vivo* experiments, many, if not all cell-based experiments of Pdx1 were performed in the context of the rat insulin promoter region, which is known to deviate from the human insulin promoter region.[7] Both *in vitro* and *in vivo* studies of the interaction were studied using the rat insulin promoter region. Nucleotide alignment of the rat and human insulin A-boxes are displayed in Table 1-2, illustrating the differences in flanking nucleotides. In addition, because the homeodomain of Pdx1 has the capability of distinguishing A-boxes *in vitro,* and shows a 5-fold

weaker binding to the IAPP promoter, it is important to draw clear connections for how Pdx1 (in the context of the full length protein) activates these promoters.

Furthermore, Pdx1 and the bHLH BETA2/E47 dimer interact synergistically to activate insulin gene expression. In the paper describing the insulin activation complex, Ohneda et. *al.* performed *in vitro*-translated interaction assays of Pdx1 and E47 to determine the regions of interaction in both protein sequences.[8] The study demonstrates interaction between the well-folded domains of Pdx1, E47, and BETA2, however, in the context of the promoter, the well-folded domains are in close proximity to warrant the interaction. Interestingly, these proteins have regions of disorder that have not been characterized that, perhaps, play a transient role in protein-protein interactions not detectable by *in vitro* interaction assays. Exploring the synergy between these proteins in the context of the human insulin and IAPP promoters and protein length would provide insight into the mechanism of gene expression, co-expression of insulin and IAPP, and flanking sequence variance in the context within the cell.

The preliminary work and future directions of this section described herein are designed provide insight into the binding of full length Pdx1 and its activation of the human insulin (hINS) and human IAPP (hIAPP) promoter sequences. Future work will illuminate whether Pdx1 plays a role in maintaining differential concentrations of insulin and IAPP. Furthermore, additional promoter sequences derived from hINS were generated using the 16 base pair consensus sequence to determine the effect of flanking sequences, spacing between core sites, and spacing between the TATA box.

6.3.2. Materials and Methods

6.3.2.1 Plasmid preparation

The pDGCR8/Pdx1 expression plasmid encodes the human Pdx1 amino acid sequence, which is controlled by the T7 promoter. First, a synthetic gene encoding human Pdx1 was purchased from GeneArt (codon optimized for mammalian cells). The gene was amplified by PCR for subsequent subcloning into the DGCR8 vector using the restriction sites, KpnI and HINDIII, to drop out the DGCR8 sequence and ligate the human Pdx1 gene. The recombinant DNA was transformed into DH5α cells for storage.

Luciferase pGL4.23 and pGL4.74 plasmids were purchased from Promega. The pGL4.23.LUC vector, used to subclone the synthetic promoter sequences, is controlled by a minimal promoter and contains the *Luciferase* gene. The *Renilla* internal control plasmid, pGL4.74.RL, is controlled by the Thymidine Kinase promoter and contains the *Renilla* luciferase gene. A synthetic polygene containing the hINS, hIAPP, and two derived promoter sequences was purchased from GeneArt. Each gene was individually subcloned into the pGL4.23 vector using the KpnI and HINDIII restriction sites (Figure 6.3-1). Sequences were aligned to the TATA box, where the first A-box was 50 and 59 bps from the TATA box, for hINS and hIAPP, respectively (Figure 6.3-1). The hINS promoter sequence was subcloned into the pGL4.23 minimal promoter to generate the pGL4.23.hINS plasmid.

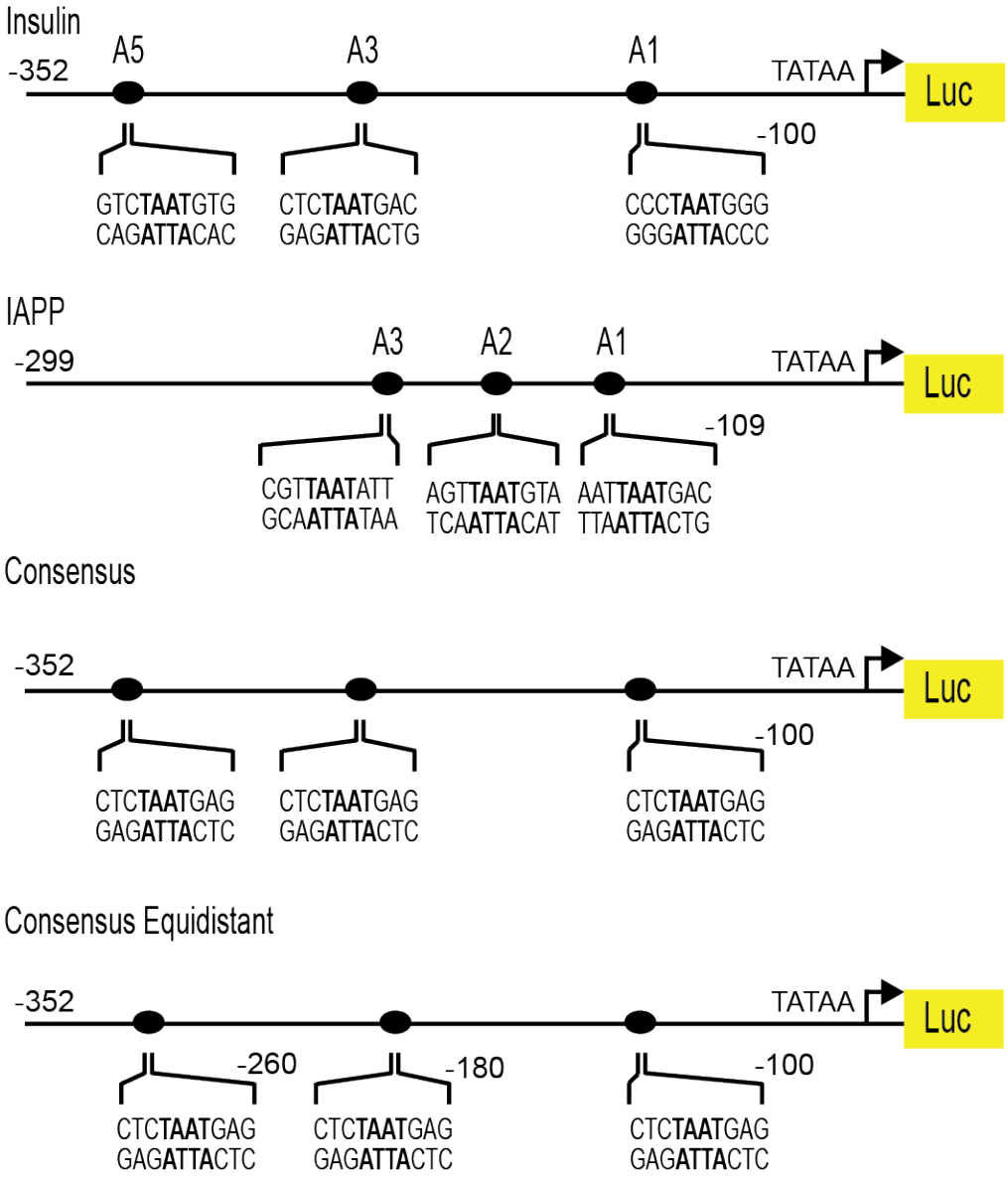

Figure **6.3-1.** Illustration of promoter region of human insulin (hINS), human IAPP (hIAPP), consensus (CONS), and consensus equidistant (CONS_EQ) luciferase plasmids. Numbering is relative to the transcription start site in the pGL4.23 plasmid.

6.3.2.2 Cell Culture

HEK 293T were cultured in DMEM media supplemented with 10% fetal bovine serum (FBS) and 1% streptomycin/ampicillin and incubated at 37 °C with 5% CO_2. Huh7.5 cells were

cultured in DMEM media supplemented with 10% FBS, 1% penicillin/streptomycin, and 1% non-essential amino acids and incubated at 37 °C with 5% CO_2

6.3.2.3 Transfections

Two days prior, cells were split in a 3:10 ratio into 6-well plates (600 µL cells, 1400 µL DMEM). Transfections of 293T cells were carried out using chemical technique with Lipofectamine 2000 (Life Technologies, Inc.). The transfection mixture was prepared by making two solutions, Solution A and Solution B. 750 ng of Pdx1, 250 ng of promoter plasmid, and 1.25 ng pGL4.74.RL were diluted in Solution A, by aliquoting 310 µL of Opti-Mem media (Invitrogen) per well. Solution B contained 310 µL of Opti-MEM and 6.2 µL Lipofectamine 2000 per well. Both solutions were incubated at room temperature for ~5 mins. 310 µL of Solution B was added to Solution A and allowed to incubated at room temperature for ~30-45 minutes. Following incubation, 1.4 mL of DMEM was added to each transfection sample (per well). The DMEM was removed from the 293T cells and the transfection mixture was added drop wise, to minimally disturb cells and to obtain adequate transfections. The cells were harvested 48 hours later. To harvest cells, the plates were taken from the incubator and placed on ice. The cells were resuspended in the DMEM media and transferred to a 15 mL conical tube on ice. Subsequently, 2 mL of cold 1X PBS was added to each well, to ensure complete resuspension of cells and collect any remaining cells, and placed on ice. To pellet the cells, the tube was centrifuged at 4200 rpm, 4 °C for 3 minutes. The media was removed and PBS buffer from the wells was added to the pelleted cells for washing. The cells were centrifuged again and the buffer was removed. For storage, the cells were flash frozen in liquid nitrogen and stored at -80 °C.

6.3.2.4 Luciferase Assay

The Dual-Luciferase Reporter assay (Promega) was used to measure luciferase activity. The manual provided with the kit was followed, except where noted. To lyse the cells, 150 µL of 1X passive lysis buffer was used to resuspend the thawed cells. The cells were allowed to incubate at room temperature for 30 mins. Luminescence was monitored using a Junior LB 9509 luminometer (Berthold Technologies).

6.3.2.5 Antibodies

Two Pdx1 antibodies were purchased to identify transfection and expression of Pdx1 within mammalian cells. The polyclonal Pdx1 antibody was purchased from Aviva Systems Biology Corp (ARP37972_P050) because the binding epitope was in a region outside the homeodomain (aa 1-60). This antibody was diluted in a 1:500 ratio for Western blot analysis. The Aviva polyclonal Pdx1 antibody was not as specific as we anticipated; several additional bands were visible. To obtain clearer results and verify Pdx1 transfection, a second Pdx1 antibody was purchased from Santa Cruz Biotechnology, Inc. The antibody was a monoclonal Pdx1 antibody (sc-398502) with the binding epitope between aa 95-110. This antibody was also diluted in a 1:500 ratio for Western blot analysis. An anti-mouse secondary antibody was purchased from Pierce (Product No. 0031430). To extend the shelf life of the secondary antibody, a final concentration of 50% glycerol was added to the mix, the solution was aliquoted and stored at -20 °C. For Western blot analysis, the secondary antibody was diluted in a 1:10,000 ratio.

6.3.2.6 Western Blots

Samples taken from chemically lysed mammalian cells were mixed with 4X protein loading dye in a 4:1 ratio. Samples were heated at 95 °C for 3-5 minutes, then 20 µL of each sample was run on a protein gel cast with 10-wells. The protein gel was electrophoresed at 190 V for 1.0-1.5 hours or until the 10 kDa marker was run off the gel. The gel was soaked in 1X transfer buffer (25 mM Tris Base, 20 mM Glycine, 0.1% SDS, 10% methanol) for 5 mins, along with 2 pieces of Whatman 3 mm paper and 1 piece of nitrocellulose membrane. The transfer unit was assembled by layering the Whatman paper, nitrocellulose, protein gel, followed by another layer of Whatman paper, with ~3 mL of 1X transfer unit added to aid in removing bubbles before each layer was placed. Excess buffer was removed and the transfer unit was allowed to electrophorese for 45-60 minutes, depending on the range of protein sizes the user wants transferred. Larger proteins > 30 kDa could take longer than 60 minutes to transfer onto the nitrocellulose. The nitrocellulose membrane was blocked with 5% w/v non-fat dry milk in TBS-T (50 mM Tris HCl, pH 7.4, 150 mM NaCl, 0.1% Tween-20) for 60 minutes at room temperature. Primary antibodies were made up in 15 mL 2% TBS-T in a 1:500 ratio for Pdx1 antibody from Aviva Systems Biology or 1:500 ratio of Pdx1 antibody from Santa Cruz Biotechnology. The membrane was incubated at 4 °C overnight in the primary antibody solution, while gently shaking. Note, antibody solutions may be reused up to one week, if stored at 4 °C. To ensure excess primary antibodies have been removed, the membrane was washed several times, the membrane was washed one for 5 minutes and twice for 10 minutes in 2% milk TBS-T. The secondary antibody was diluted in 15 mL 2% milk TBS-T at a 1:10,000 ratio for the goat anti-mouse antibody and incubated for 3 hours at room temperature, while gently shaking. The membrane was washed twice for 10 minutes with 2% milk TBS-T and twice for 5 minutes with TBS-T. For luminescence, the membrane was soaked in 1-2 mL of Lumi-Light Western Blot

substrate (SuperSignal West Dura, 34075). Equivalent volumes of the two solutions were mixed well and pipetted onto the membrane for a 5 minute incubation at room temperature, making sure to cover the surface evenly. While on saran wrap, the membrane was dried by dabbing with a clean wipe and placed on an X-ray film for exposure in a dark room.

6.3.3 Preliminary Results

6.3.3.1 in vivo DNA Binding Activity of Pdx1

A series of trials were conducted to optimize the ratios of protein to promoter and promoter to internal control. Notably, increasing the amount of Pdx1 transfected yielded surprising results. An increase in the amount of Pdx1 transfected led to a decrease in the luciferase activity. An inhibitory effect of transactivation of both insulin and IAPP promoters was observed when greater than 750 ng of Pdx1 plasmid was transfected into 60 mm^2 wells.

Transfections were completed with NIH-3T3 to no avail. However, transfections proceeded well with HEK293T and Huh7.5 cells. Surprisingly, HEK 293T cells endogenously expressed Pdx1, which was confirmed by two different antibodies (a polyclonal and monoclonal) purchased from different companies. A literature search did not result in support of endogenous expression of Pdx1in HEK 293T cells. Based on studies of the morphological development of the pancreas and duodenum, expression of Pdx1 is known to be specific to these two organs, including minor expression observed in the brain. Therefore, there was no precedence of endogenous expression of Pdx1 in HEK 293T being that they are human embryonic kidney cells.

Previous Pdx1 cell-based assays and *in vitro* binding assays were performed with DNA sequences derived from the rat promoter elements, which deviate from human promoter sequences. In our thermodynamic studies, we used sequences from the human promoters, which

produced an approximate 3-fold change in the binding affinities of IAPP versus insulin elements. To determine the effects of Pdx1 binding *in vivo*, human Pdx1 and human promoter sequences were transiently transfected into HEK 293T cells. Control experiments accounted for basal activity where only the promoter.LUC plasmid and *Renilla* plasmids were transfected. In non-control experiments the Pdx1 plasmid was also transfected to measure the transactivation activity of Pdx1. The activity was measured by computing the ratio of luciferase luminescence to renilla luminescence for both control and non-control experiments. The fold activity of non-control to control experiments is presented in Figure 6.3-2. Only a slight increase in fold activity is observed for the consensus sequence over hIAPP. The transcriptional activity of Pdx1 on the hINS_long promoter did not result in a marked increase over hIAPP. However, this may be a result of the distance of the TATA box from the first A-box element, which is not consistent with the human sequence. Although the values presented are relatively low, in the absence of binding partners these values are not uncommon in the literature.[9]

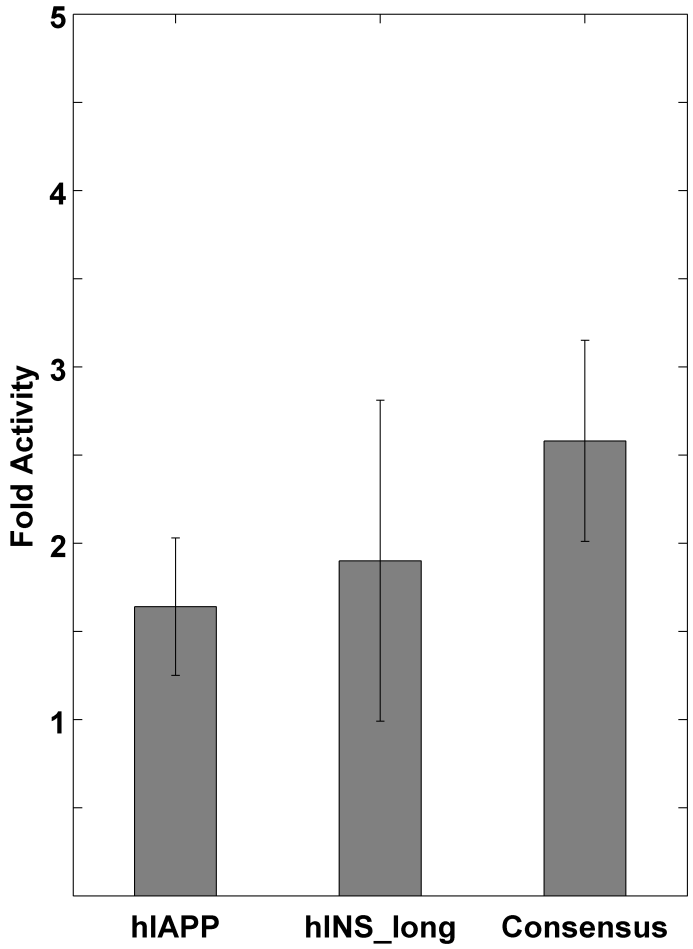

Figure **6.3-2.** Luciferase assays of transiently transfected Pdx1 with human IAPP (hIAPP), human Insulin (hINS_long), or Consensus plasmids. Fold activity is calculated by computing the ratio of the experiment versus the control. The hINS_long construct is long in the sense that the elements are not positioned correctly relative to the transcription start site as in the human promoter.

6.3.4 Discussion and Future Directions

With many of the cell-based studies of Pdx1 done using the Rat insulin promoter, the flanking sequence variation, and difference in number of A-boxes within the promoters, cell-

based studies will illuminate how Pdx1 interacts in the context of the cell. Furthermore, exploring the synergy between co-factor proteins in the context of the human insulin and IAPP promoters and protein length would provide insight into the mechanism of gene expression, co-expression of insulin and IAPP, and flanking sequence variance.

The insulin and IAPP promoters, encompassing all of the enhancer elements, and two derived consensus sequences were generated to study different aspects of this goal. In the context of the DNA-binding domain of Pdx1, I studied the binding interaction of Pdx1's homeodomain with various DNA sequences incorporating the consensus sequence derived from natural promoter elements, and insulin and IAPP promoter elements to better understand the distinct concentrations of co-secreted hormones. Chapter 2 describes in detail the findings of this study. The major finding was that the differing electrostatics of the insulin and IAPP promoter elements facilities variable binding affinity. Pdx1 homeodomain interacted preferentially to the insulin promoter elements, as compared to the IAPP elements. Thus, pursuing a more biological context of this result may provide evidence for justifying the discrepancy in insulin and IAPP secretion.

6.3.5 References

1. Westermark, P. (2011) Amyloid in the islets of Langerhans: thoughts and some historical aspects, *Upsala journal of medical sciences 116*, 81-89.
2. Westermark, P., Andersson, A., and Westermark, G. T. (2011) Islet amyloid polypeptide, islet amyloid, and diabetes mellitus, *Physiological reviews 91*, 795-826.
3. Segal, R., and Berk, A. J. (1991) Promoter activity and distance constraints of one versus two Sp1 binding sites, *The Journal of biological chemistry 266*, 20406-20411.
4. She, B. R., Schaley, J. E., and Taylor, M. W. (1995) Modulation of APRT transcription by altering spacing between cis-regulatory elements, *Somatic cell and molecular genetics 21*, 43-50.
5. Levine, M. (2010) Transcriptional enhancers in animal development and evolution, *Current biology : CB 20*, R754-763.

6. Busby, S., West, D., Lawes, M., Webster, C., Ishihama, A., and Kolb, A. (1994) Transcription activation by the Escherichia coli cyclic AMP receptor protein. Receptors bound in tandem at promoters can interact synergistically, *Journal of molecular biology 241*, 341-352.

7. Liberzon, A., Ridner, G., and Walker, M. D. (2004) Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1, *Nucleic acids research 32*, 54-64.

8. Ohneda, K., Mirmira, R. G., Wang, J., Johnson, J. D., and German, M. S. (2000) The homeodomain of PDX-1 mediates multiple protein-protein interactions in the formation of a transcriptional activation complex on the insulin promoter, *Molecular and cellular biology 20*, 900-911.

9. Matsumura, H., Kudo, T., Harada, A., Esaki, R., Suzuki, H., Kato, M., and Takahashi, S. (2007) Suppresion of MafA-dependent transcription by transforming growth factor-beta signaling, *Biochem Biophys Res Commun 364*, 151-156.

Appendix A

A User's Guide to Isothermal Titration Calorimetry

A.1 Introduction

Calorimetry is a fundamental tool that has existed for decades and is used in a variety of fields. It is used to measure the energy absorbed or released by a system. In the food science industry, calorimetry is employed to determine the calories in a particular food item, while chemists, material scientists and biophysicists use it to study the interaction between two molecules. The calorimeter used in this dissertation to study the interaction of two molecules is known as an isothermal titration calorimeter (ITC), which possesses two distinctive features (Figure A-1). First, unlike the calorimeters we learn about in classical general and physical chemistry courses, for example the coffee cup and bomb calorimeters, the ITC is engineered with a feedback circuit to maintain a constant temperature. This in turns enables the measurement of small heats with precision. Second, the ITC operates in a titration mode, either as multiple discrete titrations or a single injection titration, which is automated. In an ITC experiment, the titration of one molecule into another generates or absorbs energy in the form of heat. The increase or decrease in temperature in the sample cell creates a non-zero differential temperature between the reference and sample cells. In order to maintain the difference in temperature between the two cells at zero, power must be applied or removed, which generates the raw data for an isothermal titration experiment.

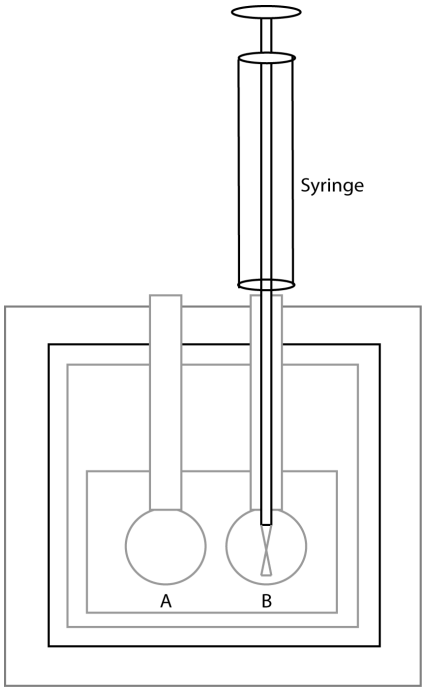

Figure **A-1**. Representation of an isothermal titration calorimeter. The two cells are shown, where A is the reference cell and B is the sample cell containing the macromolecule.

ITC is an excellent method that can be used to acquire a complete thermodynamic profile a particular system. In a single measurement the stoichiometry (n), binding constant (K_a), enthalpy of binding (ΔH), Gibbs free energy (ΔG, eq. 3-4), and entropy (ΔS, eq. 3-5) can be determined. Figure 1-6B illustrates the raw data from an ITC measurement and the integration of that data to yield thermodynamic parameters. Moreover, the heat capacity change (ΔC_p) of the system can be determined by performing a series of temperature-dependent titrations. With the heat capacity change, the temperature dependence of the typical thermodynamic parameters can be calculated. If a system interacts by displacing water molecules from hydrophobic surfaces the extent of the hydrophobic effect can be established, resulting in a large negative ΔC_p. Another important experiment, which is not as common in the literature, is the differentiation of electrostatics and non-electrostatics of a system. By studying the binding of a system at varying salt concentrations, the extent of the polyelectrolyte effect can be established. Like the

hydrophobic effect, the polyelectrolyte effect is observed when ions are displaced, which is typically observed when cations are displaced from highly anionic nucleic acid backbones.

Thus, ITC is an expedient method to determine multiple thermodynamic parameters in a single label-free experiment. However, as is the case for any method, there are some drawbacks to consider before choosing ITC. The equilibrium dissociation constant (Kd) for the interaction being measured should be significantly sub-millimolar in order to obtain acceptable quality data, clearly the interaction must generate sufficient heats, and the biomolecules must be stable at high concentrations and in the presence of its binding partner. It is, however, possible to study interactions in the millimolar range and tighter than nanomolar binding.[1, 2] To set up a successful ITC experiment, there are three critical conditions to – buffer, macromolecule and/or ligand concentration, and relative binding affinity. The outline of this appendix is as follows: considerations before setting up a titration, tips for setting up a titration, tips on the analysis of data and finally troubleshooting.

A.2 Choosing the Optimal Calorimeter

In the Chemistry department at Penn State, two isothermal calorimeter models are available. The Automated Biological Calorimetry facility houses the Auto-iTC 200 and a differential scanning calorimeter, and there are two VP-ITC instruments in the labs of Philip C. Bevilacqua Squire J. Booker and Scott A. Showalter. One thing to note, although not of much importance, is that the brand of the calorimeter, for these instruments, has changed multiple times (MicroCal, LLC to GE Technologies to Malvern), but this has little impact on calorimeter design.

Biomolecules interact across a broad range of affinities, however, not all interactions can be measured by calorimetry. The VP-ITC is best for a wide range of binding affinities and works best for studying weak interactions. First, the VP-ITC is more sensitive, meaning the signal-to-

noise is much better with this instrument than the microcalorimeter. Because the VP-ITC is more sensitive, the user can get away with lower concentrations for both the ligand and the macromolecule, so long as the interaction yields reasonable and measurable heats (> 0.1 μcal). The VP-ITC is appropriate for weaker binders because the volume of the syringe is ~10 times larger than the auto, which is advantageous for obtaining good heats by adjusting the injection volume, and providing many more points to improve fitting. It goes without saying that the VP-ITC can be used to easily measure the thermodynamics of systems binding with high affinity. There are two limitations with the VP-ITC that I have encountered. First, it is difficult to near impossible to study interactions in the high millimolar range, no matter how concentrated your ligand and macromolecule are. Second, each experiment takes upwards of 2-4 hours, before accounting for the time invested in a proper cleaning regimen. Table A-1 summarizes the general specifications required to set up an experiment. For high-throughput calorimetry, the Auto-iTC 200 is optimal.

Table **A-1.** General specification for the Auto-iTC 200 and the VP-ITC.

	Auto-iTC	**VP-ITC**
Sample Volumes:		
Syringe	120 μL	600 μL
Cell	400 μL	2100 μL
# of Injections	15-40	30-40
Injection Volume	0.5-2 μL	3-15 μL
Concentrations	higher	lower
Experimental Time with cleaning	~1.5 hrs	~3.5 hrs

The Auto-iTC-200 differs from the VP-ITC primarily in two areas. For one, a robot controls the experimental set-up and cleaning, thus this instrument is conveniently automated and excellent for high throughput. While, iTC-200 microcalorimeters are common, we at Penn State are fortunate to have an automated iTC-200 calorimeter. The volumes of the syringe and the sample cells are dramatically reduced in comparison to the VP-ITC, which can significantly decrease the number of injections achievable in a single loading. Additionally, the injection volumes are 0.5-2 μL, requiring the concentrations to be relatively high for weak binders to attain reasonable heats. The major advantage is high throughput with the ability to apply various methods within a single experimental set up. For example, the user can set up an entire temperature experiment, simply by creating the same method at different temperatures. Several disadvantages arise in choosing to use the auto calorimeter. Because the auto calorimeter provides high throughput, the user must be aware that a stringent cleaning regimen must be incorporated. Lastly, this instrument is less sensitive than the VP-ITC, meaning the signal-to-noise is lower. Therefore, if the user should wish to use the Auto-iTC 200, simulating the experiment and/or setting up a sample titration are recommended.

Briefly, to simulate the experiment on the Auto-iTC 200, the MicroCal ITC200 program should be opened under "Demo mode." The predicted stoichiometry, K_d, and binding enthalpy should be entered on the left hand side of the program as illustrated in Figure A-2. A binding isotherm will be modeled and on the right hand side of the simulated titration, the concentrations of the ligand and macromolecule and the c-value will be listed. If the c-value is out of range, it will be highlighted in red. The user can edit the concentrations of the macromolecule and ligand to obtain a modified simulated titration.

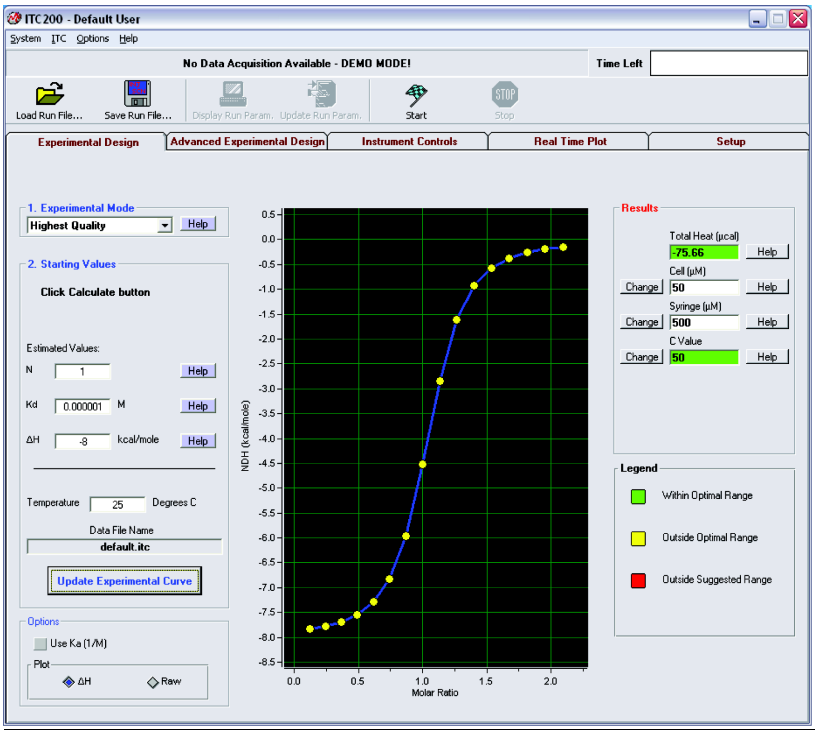

Figure **A-2.** Simulation of Auto-iTC 200 titration using the ITC200 Software.

A.3 Measures To Take Before Setting Up An Experiment

Although ITC is straightforward to run, care must be taken to ensure ITC experiments are set up optimally. With that, there are multiple conditions to prescreen to help guide the user in setting up an optimal ITC titration. The first and most important condition to test is buffer composition. Not all buffers are optimal for an ITC titration due to the possibility of measuring large heats of ionization generated from a buffer. The heat of ionization is, aptly described by the name, the energy released or absorbed by the process of ionizing the buffer. Heat of ionization of buffers could be so high that they mask the heat of binding the user wishes to observe. There are several optimal buffers for ITC (Table A-2) that have very low heats of ionization.[3] Some of the more common biological buffers, such as Tris and HEPES have extremely high heats of ionization and should be avoided, if possible (Table A-2). However, if it is the last resort, it is

possible to attain significant data using Tris and HEPES, see Tsodikov and Biswas, for example.[4] In addition, reducing agents are also used to stabilize proteins or reduce disulfide bonds, but their use in ITC is neither encouraged nor discouraged, as it can cause baseline artifacts. Importantly, the use of dithiolthreitol (DTT) should be avoided because it results in inconsistent baselines. Instead, β-mercaptoethanol or TCEP should be used at concentrations lower than 5 mM. Of particular note, the reducing agent TCEP is not stable in phosphate buffer. It has been shown that high concentrations of reducing agent produce negative effects on titrations; therefore, limiting the concentration as much as possible is always ideal.

Table **A-2.** Enthalpy for the ionization reaction of several biological buffers in water at 298 K

Buffer	pK	ΔH (kJ/mol)
Group 1: Optimal biological buffers with low heat of ionization		
Acetate	4.756	-0.41
Cacodylate	6.28	-3
Citrate	6.396	-3.38
Maleate	6.27	-3.6
Phosphate	7.198	3.6
Succinate	5.636	-0.5
Group 1: Biological buffers with high heat of ionization		
HEPES	7.564	20.4
Imidazole	6.993	36.64
Tris	8.072	47.45

Extreme measures are taken to ensure the stability of biomolecules, especially proteins. In the case of protein-protein titrations, one component may be soluble in a particular buffer, while the other may not. Or perhaps, both proteins are soluble in the same buffer, but upon titration the complex is not stable in that particular buffer. Therefore, it is highly recommended to complete a solubility study of all proteins and the varying protein constructs before trying ITC. If the proteins are soluble in the same buffer, it is wise to titrate one protein into the other to

simulate an ITC experiment in a microcentrifuge tube to avoid clogging up the calorimeter (cleaning will be discussed below).

Concentration is a critical parameter that must be quantitatively determined for setting up an optimal ITC titration. The concentrations of both components should be known accurately as data analysis requires the concentration for fitting the data. Stoichiometry and, in some cases, the apparent binding affinity are affected by incorrect concentrations; therefore, knowing the concentration is imperative. Determining biomolecule concentrations by UV spectroscopy is highly recommended. Nevertheless, there are proteins and peptides that lack aromatic residues; as such their concentration cannot be determined by UV spectroscopy. Other methods such as FT-IR, quantitative protein gels or BCA/Bradford assays can be used to determine the concentration, albeit with greater error.

Once the concentrations of the components are known, the relative binding affinity can be qualitatively determined, if the user has no reference value. Before starting a titration, the user should become familiar with the sample preparation and experimental set-up, described below. There are several possibilities for the relative binding affinity regime, femtomolar to sub-nanomolar, nanomolar to micromolar, and micromolar to millimolar. Before going into the typical shape of the sigmoid from the binding thermogram, the Wiseman coefficient (or c-value) will be discussed.

The Wiseman coefficient, referred to as the c-value from here on out, is described by the following equation

$$\boldsymbol{c = nK_a[M]_{tot}} \tag{A.6}$$

where c is the Wiseman coefficient, n is the stoichiometry, and M is the total macromolecule concentration. Depending on the regime of the K_a, the c-value is used to predict and determine the shape of the binding sigmoid. For c-values ≤ 25, the sigmoid will be very shallow, indicative of

weak binding in the micromolar to millimolar range, typically with little to no baselines. The optimal c-value, which is still debated in the literature, falls within 50-500, which results in a complete sigmoidal curve, usually with an upper and lower baseline. C-values greater than 500, typically those closer to 1000, will yield a sigmoid that can be described by a step function. Clearly, there are advantages and disadvantages to collecting ITC experiments where the c-value falls outside the 50-500 range.

For systems that interact with weak binding affinity, the c-values are typically below 10. Binding thermograms for these systems usually lack one or both baselines, which results in erroneous binding enthalpies. However, the stoichiometry and the binding affinity, if the concentrations are accurate, can be fit with good statistics. Moreover, the thermogram for interactions where $c < 1$ are often difficult to fit to yield valuable data.[5]

On the other extreme, the c-value for extremely tight binders in femptomolar to subnanomolar regime is greater than 500, typically upwards of 1000. As a result, the thermogram yields a stepwise function, with no points sampling the slope of the thermogram. In this case, the binding enthalpy and stoichiometry is determined accurately, while the binding affinity is estimated. For extremely tight binders, the user can opt to set up a competition assay to quantitatively determine the binding affinity of the system, an excellent methodological article by Krainer et *al.*[6] can be reviewed. For the extremes, very low ($K_a < 10^4$ M^{-1}) and very high ($K_a > 10^8$ M^{-1}) affinity interactions, competition or displacement experiments can be completed to determine binding affinity and/or binding enthalpy accurately. The reader is referred to several excellent articles for competition or displacement assays using ITC.[6-8]

Lastly in regards to the c-value, the optimal range for attaining quantitative fits of the stoichiometry, binding affinity and binding enthalpy is 50-500. If the user is setting up a titration where the projection of the c-value is in the upper limit, sampling the midpoint of the thermogram will result in fewer points. A word of advice for the user is to maintain a happy place for the

protein(s) being studied. The following section will introduce the controls prior to completing a series of titrations.

A.4 Tips For Titration Set Up

For each system tested, whether a new protein is being tested or simply a different construct, the measures in the previous section should be retained. Equally important to concentration and buffer composition is the requisite of setting up proper controls to yield faithful data. Therefore, at least 3 controls should be run before continuing with the completion of the study: 1) buffer into buffer, 2) ligand into buffer, and 3) buffer into macromolecule. The first, buffer into buffer, will establish two things – complete dialysis or identical buffer compositions and the heat of ionization of the buffer. The second, ligand into buffer, will control for any secondary processes that may be occurring absent binding, such as concentration-dependent oligomerization. Similar to titrating ligand into buffer, the control titration of buffer into macromolecular will establish the (hopefully) absent additional processes. In addition, if no upper baseline is obtain from the experimental titrations, the second control can be used as a baseline subtraction for subsequent experiments completed in the same conditions (i.e., buffer, concentration and temperature).

Before the user begins the experimental set-up, several sample conditions and instrument conditions must be met. As mentioned above, the ligand and macromolecule must be in identical buffer compositions to minimize heats of dilutions. Because ITC is a sensitive technique, small inconsistencies in buffer composition can yield significant heats of dilution that can mask or ruin a titration. Thus, overnight co-dialysis of the biomolecules is extremely important. Before dialysis, typically one concentrates and buffer-exchanges the samples using centrifugal concentrators (we find that Millipore PES spin columns perform well for proteins) prior to co-

dialysis. Of particular note, when using any type of spin column, the user should be aware that the membranes are coated with glycerol or PEG to maintain the membrane during long storage. The column should be rinsed multiple times and the user is encouraged to centrifuge ~20-40 mL of water or buffer through the membrane prior to using it for biomolecule concentration. Lastly, it is imperative that the dialysis buffer be filtered prior to running an experiment.

Although the instrument should be maintained and cleaned following the completion of experiments, it is in the best interest of the user to monitor the condition of the instrument. To obtain useful data, the cleanliness of the instrument is of utmost importance. To determine the cleanliness and efficiency of the calorimeter, the user should run several controls after completing the cleaning regimen. Specifically, a water-water titration will report on the cleanliness of the cell and in some cases the syringe. Similarly, an EDTA-Ca titration can be performed to report on the cleanliness and efficiency of the calorimeter. Typically, if the instrument will be left idle for a short period of time, the sample cell should be stored in water; however, if the instrument will be left idle for a prolonged period of time, the sample cell should be store dry.

A.5 Experimental Set Up: Best Practices

In an ITC experiment, bubbles should always be avoided because they cause artifacts in the titration. There are a couple considerations and practices to alleviate bubbles from poisoning a titration. Bubbles can arise when the cell or syringe is being loaded, from inadequate degassing of the samples, or from a dirty cell. In this section, several practices are outlined for the user.

One way to avoid bubbles is to remove them from samples by degassing. Each calorimeter is supplied with a degasser, which can be held at a constant temperature. The samples should be degassed at the temperature of the experiment or 2-3 degrees below for at least 5 minutes for samples being stirred and 10-15 minutes for unstirred samples. Another way to avoid

introducing bubbles is to practice loading samples. It is extremely easy to introduce bubbles into proteinaceous samples, so I have described the best method for loading the cell without bubbles below.

The user should become familiar with the long stem Hamilton syringe to load the sample and practice their loading technique. Because samples are valuable, the user will ensure all solution has entered the barrel of the syringe, by slowly pulling on the plunger and watching for a bubble to form in the barrel. This ensures the entirety of the solution is in the barrel and minimizes sample loss. Next, to remove the bubbles, first try to center the bubble to the opening, then slowly push the bubble out making sure it stays centered with the opening by tilting, if needed. Once you can no longer see the bubble, continue to slowly press on the plunger, while observing the bottom of the Hamilton syringe until you see a drop of the solution form or fall. You have successfully removed the bubbles and can continue with loading the cell. I recommend an untrained user to begin with setting up a water-water titration to practice setting up titrations.

When loading the syringe, the user should take advantage of the purge and refill options in the parameters section. A plastic syringe with a hose attached is supplied with the instrument to load the ITC syringe. To avoid contaminating the hose and possibly the sample, try to keep an eye on the syringe barrel and the sample tube. Always load the syringe slowly and carefully as to not introduce air into the syringe. The user can see (and control) the level of the syringe. Once the solution is seen, keep an eye on the sample volume in the loading tube. When the syringe is full, take the mouse and click on close fill port. Loading the syringe requires multiple steps and also takes practice. Once loading the syringe has been completed, the user should always 'purge and refill' the syringe solution to remove any bubbles and also mixes the solution to ensure homogenous sample concentration.

Aside from setting up an experiment properly, the reference cell should be filled with degassed or filtered water at least once or twice a month, if the instrument is used regularly. Be

sure to change the reference cell before starting a titration, if the instrument was sitting idle for a prolonged period of time. The level of the reference cell should be filled flush with the base of the neck of the cell. An under filled or overfilled reference cell can effect the titration.

A5.1 Experimental Parameters

ITC parameters:

1) **Total Injections** – This is dependent on the volume of each titration, but typically on the order of 30-40 injections. The user can overestimate this value, without adversely effecting the titration; the instrument will report insufficient volume in the syringe and save the data. In addition, the first injection is always set to a lower volume to account of inconsistencies in syringe filling and diffusion, but this point is not included in the analysis.[8-10]
2) **Cell Temperature** – The temperature can be set anywhere from 2 to 80 °C, but the typical range is 5 to 40 °C.
3) **Reference Power –** The reference power is the constant power applied to equilibrate the temperatures between the reference and sample cells. Typically the value is set to 5, but for interactions yielding large heats, the reference power should be set to 10 or greater.
4) **Initial Delay** – Set to the minimum value of 60 seconds to establish a baseline.
5) **Concentrations** – The concentrations of the samples should be entered, but if necessary, minor corrections can be made during analysis.
6) **Injection parameters**
    a. **Volume** – The volume of each injection.
    b. **Duration** – The duration in seconds of each injection, which is defaulted to twice the value of the volume.

c. **Spacing** – Each injection should return to equilibrium or baseline. The spacing between each injection will need to be determined based on the system. Typically, 240 or 360 seconds is a good starting point.

d. **Filter Period** – This parameter alters the period in which a single data point is collected. For reactions occurring rapidly, 2 seconds is sufficient to represent a peak, which can be properly integrated to produce an accurate enthalpy.

7) **Feedback Mode/Gain** – The feedback mode dictates the response time of the instrument. For increased sensitivity or monitoring kinetics or rates, the passive mode will provide the best data. For the quickest response time, a high response time is selected. This mode is recommended in the manual.

8) **ITC Equilibration Options** – Typically, the last two (Fast Equil. and Auto) are selected. This will ensure the temperature of the cell and sample are within the starting temperature and complete an automatic period of equilibration with stirring and final baseline equilibration.

9) **Tray Temperature** – This is only for the Auto-iTC. The tray temperature can be adjusted to maintain a slightly colder temperature if stability of the samples is compromised at room temperature. If the titration temperature is set to room temperature, it is recommended to set the tray temperature to no lower than 15 °C. Anything lower can introduce air bubbles into the titration and ruin an experiment. This is due to the enhanced solubility of gas at lower temperatures.

A5.2 Data Analysis

There are several programs to analyze and fit ITC data – Origin, SEDPHAT, Affinimeter, or NanoAnalyze. Origin is provided with the instrument to process the ITC data files and is

relatively user friendly. The following discussion will be on data analysis of VP-ITC and Auto-iTC data using Origin. The analysis is the same for both instruments, with one exception for the Auto-iTC. When Origin software is started on the computer controlling the Auto-iTC, a small window opens with choices: Auto-iTC 200, ITC 200, following a few others. It is important to select "ITC 200" to analyze the data. Now, when Origin is running the user should load the data. Before any adjustments or fitting of the data, the user should adjust the concentrations, if necessary. It is advised to always verify the peak integration limits; often it is best to adjust the integration of all the peaks to maintain consistency in data analysis. At this point, to avoid any inconveniences, the project should be saved. A baseline is subtracted either from a ligand into buffer titration or average of the upper baseline points. Regarding the latter, interactions in which more than five points exist in a saturated upper baseline, the user can determine the mean of 5-10 points. To execute this step, the user will go to Window → 'FileName' → select the last 5-10 points in the 'NDH' column, right click, and select 'statistics on column.' A window will pop up displaying several parameters; copy the 'mean value;' this value will be subtracted using the 'Simple Math' function. First, return to the 'DeltaH' window. To get to the simple math menu, click on Math → Simple Math. Once the baseline has been subtracted from the data, the data should have been shifted closer to 0. Be sure that the data DOES NOT cross zero, otherwise Origin will not be able to fit the data. If the data is not up against zero in the positive range or negative range, the data will be fit properly, but inconsistently. Now, select one of the fitting methods and a new window will pop up, select '100 iter' until the Chi-square value is optimized. A final figure can be obtained from ITC → Final Figure.

A5.3 Cleaning

It cannot be stressed enough that maintaining a clean ITC is imperative! Biomolecules, especially proteins, which are sticky can quickly leave a cell or syringe dirty. Thus, the users data suffers from the perils of a dirty calorimeter. Depending on the system studied, the user must be aware of how many experiments can be run to be efficient, yet maintain a clean instrument. A good rule of thumb is to clean after 3 titrations involving biomolecules. There are three different cleaning regimens to follow depending on the situation. For typical cleaning after each titration, the user should run 50-100 mL of 10 % Contrad 70 (a strong detergent) through the syringe at room temperature, followed by at least 200-400 mL of ddH_2O through the syringe and > 600 L through the cell. After multiple experiments, even after following the typical cleaning regimen, the instrument will need a harsher clean. It is recommended that ~ 5 mg/mL of Proteinase K be incubated in the cell and syringe at 37 °C overnight, followed by a contrad cleaning for one hour at 60 °C (no more than two hours). The user should add the enzyme solution to the syringe and insert it into the cell to incubate at 37 °C. Proteinase K can be resuspended in water, or for increased activity at 37 °C, the buffer should be 10 mM Tris, 1-5 mM $CaCl_2$, pH 8.0. It goes without saying that the cell and syringe should be rinsed thoroughly with >1 L of ddH_2O. Lastly, if the previous step is not enough to produce a clean instrument, a more stringent protocol should be followed. Instead of incubating 10% contrad in the cell for 1 hour at 60 °C, the cell should be cleaned for 2 hours, followed by a quick rinse with methanol through the syringe and cell. Of particular note, when contrad is in the cell at 60 °C, the solution should be removed ONLY when the cell has cooled to near room temperature.

A.6 Troubleshooting ITC

Once the experimental parameters have been optimized for a particular system, the majority of the issues with ITC arise due to a dirty cell or syringe. Even though a good cleaning regimen is practiced after each titration with biomolecules, the instrument can still produce characteristics of a dirty cell or syringe. A low or high baseline (±1 μcal/sec from the reference power) is indicative of a dirty cell. However, a low baseline could also be indicative of an under filled cell, which could be a result of a dirty cell trapping an air bubble or insufficient transfer to the cell. Similarly, if the reference cell is under filled the baseline will be higher. To ensure the cell is clean, the cleaning regimen described above should be followed.

There are a couple issues that may arise that are out of the control of the user, such as drifts in the baseline due to temperature changes within the environment of the instrument and issues with the electronics of the instrument. First, temperature drifts in the lab impact the baseline of the experiment; these drifts should not impact the experiment, but the user should ensure that each peak is properly integrated. Occasionally, when the instrument has been left on for too long, the experiment can result in erratic thermograms (Figure A-3). If this occurs, the user should close the programs and turn the instrument off for a short period of time. Turning off the instrument and restarting the programs should remedy the problem.

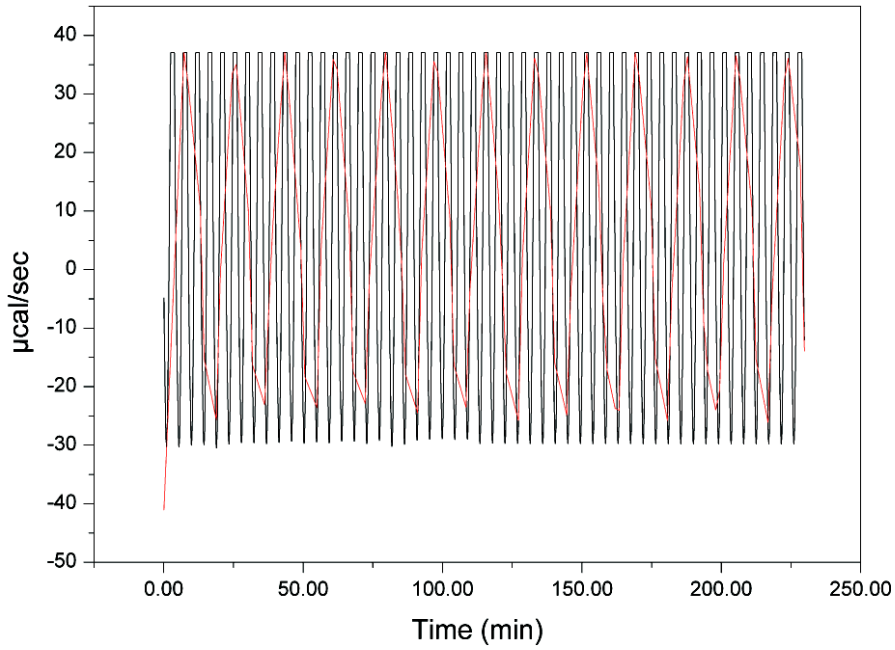

Figure **A-3.** Representative titration resulting from instrument failure.

A.7 References

1. Kabiri, M., and Unsworth, L. D. (2014) Application of isothermal titration calorimetry for characterizing thermodynamic parameters of biomolecular interactions: peptide self-assembly and protein adsorption case studies, *Biomacromolecules 15*, 3463-3473.

2. Velazquez Campoy, A., and Freire, E. (2005) ITC in the post-genomic era...? Priceless, *Biophysical chemistry 115*, 115-124.

3. Goldberg, R. N. K., N.; Lennen, R. M. (2002) Thermodynamic Quantities for the Ionization Reactions of Buffers, *Journal of Physical and Chemical Reference Data 31*.

4. Tsodikov, O. V., and Biswas, T. (2011) Structural and thermodynamic signatures of DNA recognition by Mycobacterium tuberculosis DnaA, *Journal of molecular biology 410*, 461-476.

5. Turnbull, W. B., and Daranas, A. H. (2003) On the value of c: can low affinity systems be studied by isothermal titration calorimetry?, *Journal of the American Chemical Society 125*, 14859-14866.

6. Krainer, G., Broecker, J., Vargas, C., Fanghanel, J., and Keller, S. (2012) Quantifying high-affinity binding of hydrophobic ligands by isothermal titration calorimetry, *Analytical chemistry 84*, 10715-10722.

7. Sigurskjold, B. W. (2000) Exact analysis of competition ligand binding by displacement isothermal titration calorimetry, *Analytical biochemistry 277*, 260-266.

8. Velazquez-Campoy, A., and Freire, E. (2006) Isothermal titration calorimetry to determine association constants for high-affinity ligands, *Nature protocols 1*, 186-191.

9. Ikenoue, T., Lee, Y. H., Kardos, J., Yagi, H., Ikegami, T., Naiki, H., and Goto, Y. (2014) Heat of supersaturation-limited amyloid burst directly monitored by isothermal titration calorimetry, *Proceedings of the National Academy of Sciences of the United States of America 111*, 6654-6659.

10. Pierce, M. M., Raman, C. S., and Nall, B. T. (1999) Isothermal titration calorimetry of protein-protein interactions, *Methods 19*, 213-221.

VITA

Monique Bastidas

EDUCATION

Ph.D. Chemistry
08/2015
The Pennsylvania State University

BS. Biochemistry
05/2010
California State University, Sacramento

PUBLICATIONS

Bastidas, M., Gibbs, E.B., Sahu, D., and Showalter, S.A. (2015) "A Primer for Carbon-Detected NMR Applications to Intrinsically Disordered Proteins in Solution." *Concepts Magn. Reson. Part A,* in press.

Sahu, D. **Bastidas, M.**, and Showalter, S.A. (2014) "Generating NMR Chemical Shift Assignments of Intrinsically Disordered Proteins Using Carbon-Detect NMR Methods." *Anal. Biochem.,* **449**, 17-25.

Bastidas, M., and Showalter, S.A. (2013) "Thermodynamic and Structural Determinants of Differential Pdx1 Binding to Elements from the Insulin and IAPP Promoters." *J. Mol. Biol.*, **425**, 3360-3377.

PRESENTATIONS

American Society of Biochemistry and Molecular Biology, *Boston, MA*
"Probing Disordered Region Influence on Pdx1 Binding to Natural Promoters and Near-Consensus Sites" (Poster Presentation)

Gordon Research Conference: Intrinsically Disordered Proteins, *Easton, MA*
"Thermodynamic and Structural Determinants of Differential Pdx1 Binding to Elements from the Insulin and IAPP Promoters" Poster Presentation)

27th Annual Gibbs Conference on Biothermodynamics, *Carbondale, IL*
"Thermodynamic and Structural Determinants of Differential Pdx1 Binding to Elements from the Insulin and IAPP Promoters" (Poster Presentation)