RESEARCH ARTICLE

PROTEINS WILEY

Structural and functional analysis of the human cone-rod homeobox transcription factor

Penelope-Marie B. Clanor[1] | Christine N. Buchholz[2] | Jonathan E. Hayes[2] | Michael A. Friedman[1] | Andrew M. White[2] | Ray A. Enke[1,3] | Christopher E. Berndsen[2,3] 🆔

[1]Department of Biology, James Madison University, Harrisonburg, Virginia, USA

[2]Department of Chemistry and Biochemistry, James Madison University, Harrisonburg, Virginia, USA

[3]Center for Genome and Metagenome Studies, James Madison University, Harrisonburg, Virginia, USA

Correspondence
Christopher E. Berndsen, James Madison University, Department of Chemistry and Biochemistry, 901 Carrier Drive, MSC 4501, Harrisonburg VA 22807, USA.
Email: berndsce@jmu.edu

Abstract

The cone-rod homeobox (CRX) protein is a critical K50 homeodomain transcription factor responsible for the differentiation and maintenance of photoreceptor neurons in the vertebrate retina. Mutant alleles in the human gene encoding CRX result in a variety of distinct blinding retinopathies, including retinitis pigmentosa, cone-rod dystrophy, and Leber congenital amaurosis. Despite the success of using in vitro biochemistry, animal models, and genomics approaches to study this clinically relevant transcription factor over the past 25 years since its initial characterization, there are no high-resolution structures in the published literature for the CRX protein. In this study, we use bioinformatic approaches and small-angle X-ray scattering (SAXS) structural analysis to further understand the biochemical complexity of the human CRX homeodomain (CRX-HD). We find that the CRX-HD is a compact, globular monomer in solution that can specifically bind functional cis-regulatory elements encoded upstream of retina-specific genes. This study presents the first structural analysis of CRX, paving the way for a new approach to studying the biochemistry of this protein and its disease-causing mutant protein variants.

KEYWORDS
CRX, homeodomain, photoreceptor, small-angle X-ray scattering

1 | INTRODUCTION

The vertebrate retina is a multilayered neural sensory tissue located in the posterior portion of the eye. Visual perception occurs when photons of light enter the front of the eye, are focused onto the retina, and are then absorbed by rod and cone photoreceptors, initiating the biochemical process of phototransduction.[1] Several retina-specific transcription factors (TFs) regulate the development, differentiation, and maintenance of these critical photoreceptor neurons.[2] One of these TFs, *cone rod homeobox* (CRX) has been identified as a key regulator of photoreceptor transcriptional networks during differentiation and into cellular adulthood.[2]

TFs directly affect the differentiation and maintenance of diverse cell types in various tissues by controlling transcriptional networks. Homeodomain TFs contain a homeodomain (HD), a globular domain of about 60 amino acids conferring the biochemical ability to bind specific DNA motifs.[3] HD family proteins represent 15%–30% of TFs in plants and animals and regulate diverse developmental processes.[4] Paired-type HD TFs recognize a TAAT core DNA motif which is further specified by the amino acid at position 50 of the HD.[3] This sequence

---

Marie B. Clanor and Christine Buchholz contributed equally as co-first authors.

specificity can be binned into distinct subclasses. HDs with a glutamine residue at position 50 (Q50 type) preferentially bind a TAATTA/G-motif, while those with a lysine residue at position 50 (K50 type) favor a TAATCC motif.[5–7] CRX is a critical K50 HD TF for regulating photoreceptor and bipolar cell transcriptional networks during differentiation and into cellular maturity in the vertebrate retina.[2,8] CRX functions by binding to TAATC cis-regulatory elements (CREs) encoded in genomic DNA sequences in and adjacent to photoreceptor-specific genes regulating their transcription.[9,10] CRX binding to CREs is mediated by the highly conserved N-terminal HD DNA binding domain.[11] Transcriptional regulation is achieved by an array of protein–protein interactions mediated by CRX's activation domain (AD) encompassing the C-terminal half of the protein (Figure 1A).[11,12]

Numerous clinically relevant mutations in the *CRX* gene result in an array of biochemical dysfunction and they are responsible for multiple blinding retinopathies, including retinitis pigmentosa (RP), cone-rod dystrophy (CoRD), and Leber congenital amaurosis (LCA).[13] Due to the complexity of the CRX protein, these mutant alleles can be categorized into subclasses based on allele type, biochemical dysfunction

of the protein, and resulting pathology.[14] While animal models and in vitro experiments have helped characterize the various mechanisms of CRX-induced pathology, there are currently no high-resolution NMR or X-ray structures in the published literature for the CRX protein to aid in this endeavor. In this study, we use small-angle X-ray scattering (SAXS) to analyze the three-dimensional properties of the human CRX DNA binding HD protein. This study presents the first structural analysis of this clinically relevant protein and opens the door to a new modality for studying CRX and its disease-causing mutant alleles.

2 | METHODS

2.1 | Reagents and peptides

All buffers and chemicals were ordered from Fisher Life Science or Sigma-Aldrich. A peptide corresponding to the human CRX homeodomain (amino acids 39–98) was synthesized by Genscript.
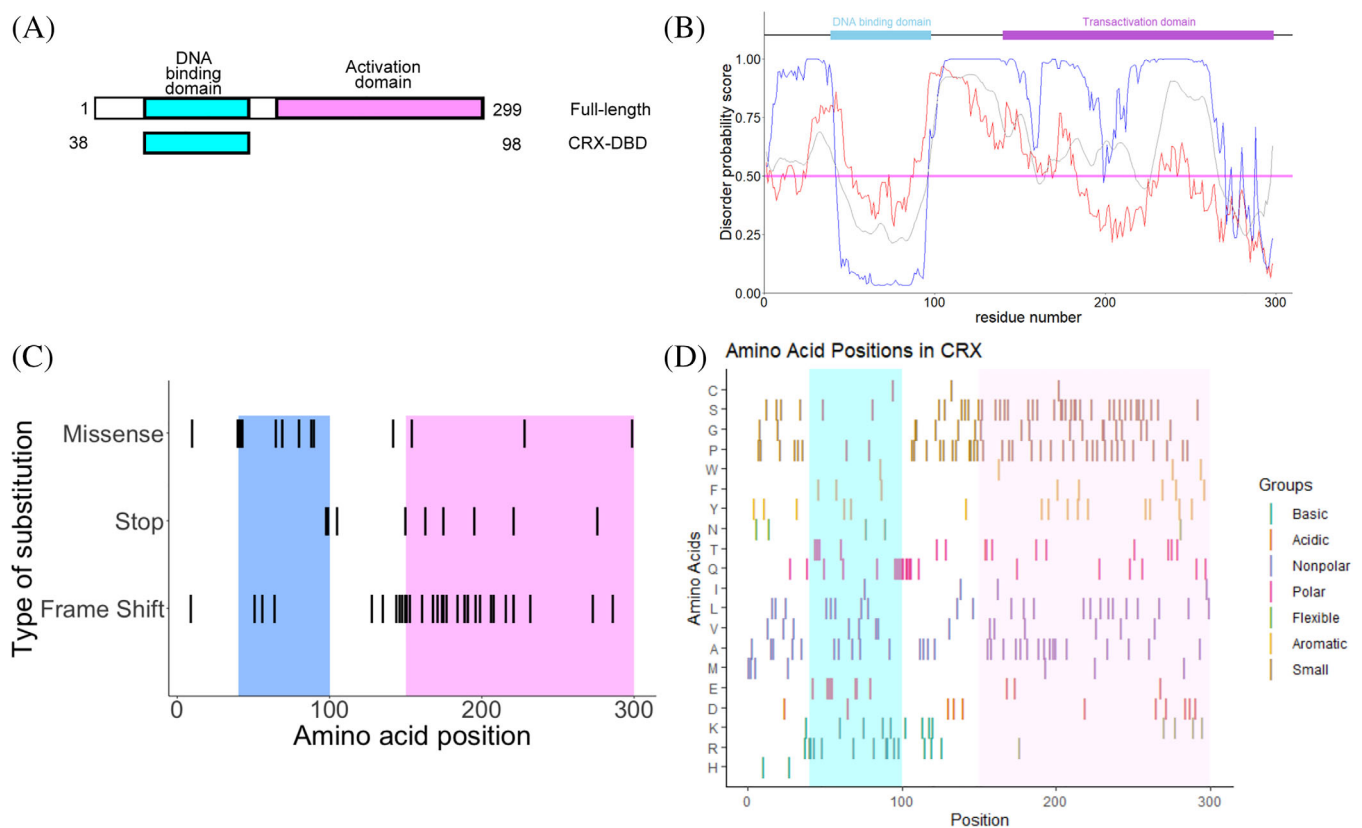
FIGURE 1 Human CRX protein organization and sequence analysis. (A) Cartoon showing the domain structure of human CRX (referred to as full-length CRX) and the DNA-binding domain of CRX (CRX-DBD). Numbers indicate the amino acids positions encompassed in each construct. (B) Disorder prediction of the full-length CRX sequence. The data from IUPred2A shown in red, from ODIN shown in blue, and from PrDOS in gray (see methods for details). (C) Distribution of clinically relevant mutations in the human CRX protein. Stop, missense, and frameshift mutations in human CRX protein were obtained from the ClinVar database. The distribution of mutations was plotted by amino acid position in the human CRX protein with cyan indicating the DNA-binding domain and salmon box indicating the activation domain. The DBD contains 13 pathogenic missense mutations and three pathogenic frameshift mutations. The activation domain contains three pathogenic missense mutations and 25 pathogenic frameshift mutations. (D) Amino acid bias within the domains of CRX. The cyan box indicates the DNA-binding domain and the salmon box indicates the activation domain

2.2 | Computational sequence analysis

DNA and protein sequences of human CRX were obtained from Uniprot entry O43186. For the sequence properties analysis in Figure 1B and D, the protein sequence was imported into R, and the amino acids were binned by properties. For visualization of the pathogenic substitutions, data were downloaded from a search of CRX from ClinVar.[15] In R, substitutions linked solely to CRX and which were described as pathogenic or likely pathogenic and either missense, premature stop, frameshift substitutions were identified. Data were then plotted using the ggplot2 package.

The primary sequence of CRX was provided to PrDOS,[16] IUPred2A,[17] and ODiNPred[18] servers for disorder prediction. The resulting predictions were plotted using the ggplot2 package.

CRX binding regions (CBRs) obtained from a previously published human retina ChIP-seq study[10] were visualized in the UCSC Genome Browser, hg38 human genome assembly, and were used to identify putative CRX binding motifs in the proximal promoter regions of the human *RHO* and *PDE6B* genes.

2.3 | CRX cloning and expression

The coding sequence for the N-terminal 107 amino acids comprising the DNA binding domain (DBD) of the human CRX protein was fused to a 6x His-tag inserted following the met start codon and subcloned into the commercial PMAL-c5x expression plasmid containing a maltose binding domain (MBD) coding sequence using Xmnl and EcoR1 restriction sites (GenScript, Piscataway, NJ). CRX DBD-MBD plasmid was transformed into BL21 *Escherichia coli* competent cells and grown on LB plates supplemented with ampicillin for selection overnight at 37°C. Transformant colonies were used to inoculate 10 ml starter cultures followed by 1 L LB cultures grown to an optical density of 0.5–0.7 OD at 600 nm at which time they were induced with 0.2 mM IPTG and cultured at 37°C for 1 h at 200 RPM. Cultures were pelleted, resuspended in lysis buffer, and disrupted using three rounds of probe sonication at 75% amplitude for 5 s on/off pulses with 1 min between pulse sequences. Lysates were pelleted and the soluble fraction was used for downstream protein purification steps.

2.4 | Protein purification

The 6X-His-CRX-DBD-MBD fusion protein was purified from lysates using a multistep chromatography strategy on an AKTA Start FPLC. Initial purification of the MBD portion of the fusion protein was facilitated using amylose chromatography. The 1 ml MBPTrap™ HP column was equilibrated with 20 mM Tris–HCl, 200 mM NaCl, 1 mM EDTA (pH 7.4) binding buffer and protein was eluted with binding buffer with an additional 10 mM maltose. Fractions containing CRX-DBD-MBD, identified by SDS-PAGE and FPLC UV spectra, were pooled. Heparin column chromatography was used to facilitate the removal of contaminating bacterial nucleic acids from the fusion protein. The 1 ml HiTrap™ column was equilibrated with 0.1 M sodium phosphate

(pH 7.4) and protein eluted with 10 mM sodium phosphate, 2 M NaCl using a linear gradient yielding 72 column volume (pH ~7). Size exclusion chromatography was used as a final step to separate heavier and lighter molecular weight contaminating proteins from the desired protein. The Hiload™ 16/600 Superdex™ 200 pg column was equilibrated with 0.02 M NaPO₄, 50 mM NaCl (pH 7). All FPLC fractions were collected and examined with SDS PAGE and 260 nm/280 nm ratio to determine purity, quality, and quantity of the final purification product.

2.5 | Small angle X-ray scattering (SAXS) analysis

The CRX DBD peptide was resuspended in water before being dialyzed overnight at 4°C into 50 mM sodium phosphate, pH 7, 100 mM NaCl, and 5 mM imidazole. A CRX-DBD concentration of 7 mg/ml was measured using the predicted extinction coefficient of 8480 M^{-1} cm^{-1} calculated from the sequence in ProtParam.[19] Samples were shipped overnight at 4°C to the SIBYLS beamline at the Advanced Light Source. Samples were injected into an Agilent 1260 series HPLC with a Shodex KW-802.5 analytical column at a flow rate of 0.5 ml/min. Small-angle X-ray scattering (SAXS) data were collected on the elution as it came off of the column. The incident light wavelength was 1.03 Å at a sample to detector distance of 1.5 m. This setup results in scattering vectors, q, ranging from 0.013 to 0.5 $Å^{-1}$, where the scattering vector is defined as $q = 4\pi\sin\theta/\lambda$ and 2θ is the measured scattering angle. Data collection statistics are available in Table 1.

TABLE 1 Small Angle X-ray Scattering (SAXS) data collection and analysis parameters

Instrument	SIBYLS beamline 12.3.1
Wavelength (Å)	1.03
q range ($Å^{-1}$)	0.027–0.46
Injected Concentration (mg/ml)	7
Temperature (K)	283
Column	Shodex KW-802.5
I (0) (A.U.) [from p (r)]	0.93 ± 0.01
R_g (Å) [from p (r)]	12.4 ± 0.1
I (0) (A.U.) [from Guinier]	0.91 ± 0.01
R_g (Å) [from Guinier]	11.8 ± 0.2
D max (Å)	40.0
Porod volume estimate ($Å^3$)	3670
Molecular mass Bayes (Da)	5700 [99.6% C.I. 5.0–6.0]
Calculated monomeric Mr from the sequence (Da)	7369
Primary data reduction	Scatter 4.0d and RAW 2.1.1
Data processing	RAW 2.1.1
Dummy atom modeling	DAMMIF
Electron density calculation	DENSS
Three-dimensional graphic representations	Chimera

Radially averaged data were processed in SCATTER (ver 4.0d) and RAW (ver. 2.1.1) to remove the scattering of the sample solvent and identify peak scattering frames.[20] RAW was used to calculate the dimensions of the molecule, the molecular weight, and the pair-distance distribution function (PDDF). Plots of data were produced in R using the ggplot2 package.[21] The SAXS intensity plots for Drosophila Pax homeodomain (1FJL),[22] Drosophila Aristaless (3A01),[23] and human PAX3 (3CMY)[24] using the FOXS web server.[25] Predicted data were overlaid on the data for the CRX-DBD in R.

2.6 | Ab initio modeling of CRX shape

The electron density of the peptide was calculated in RAW using the DENSS program.[26] DENSS predicted 51 shapes followed by averaging and refinement. The SAXS data were also used in the DAMMIF program in RAW to create a dummy atom model.[27] The DAMMIF program created 15 models followed by refinement and averaging. The results from DENSS and DAMMIF were each visualized in the program CHIMERA to produce a mesh model from each.[28] Each of these models were then individually fitted with a homology model created from the DBD sequence in the program trRosetta.[29]

2.7 | Circular dichroism

For CD measurements, the cuvette contained CRX-DBD at a concentration of 24 μM in 50 mM sodium phosphate, 100 mM sodium chloride, and 5 mM imidazole in a 0.2 cm quartz cuvette. Data were collected from 320 to 200 nm at 1 nm/s with three accumulation on a Jasco J-1500 CD spectrometer. Spectra were collected at temperatures from 10–95°C at an interval of 5°C, with a gradient of 2°/min. The sample was equilibrated at each temperature for 1 min before spectra were collected. Data were exported in .csv formatted and plotted in R using the ggplot2 package. Apparent melting temperatures were calculated through a global fit to a one-step equilibrium unfolding model using the Calfitter app.[30] CD spectra were predicted for the Drosophila Pax homeodomain (1FJL),[22] Xenopus TFIIIA zinc finger (1TF6),[31] mouse OTX2 (2DMS), Caenorhabditis CEH37 (2MGQ), Drosophila Aristaless (3A01),[23] human PAX3 (3CMY),[24] and mouse p53 core (3EXJ)[32] using PDBMD2CD.[33] The predicted spectra were plotted in R and overlaid on the data for the CRX-DBD peptide. The predicted data were normalized to the 222 nm value for the CRX-DBD peptide.

2.8 | DNA binding assays

Thity five base pair double-stranded DNA oligonucleotide probes corresponding to putative CRX binding sites within the human *RHO* and *PDE6B* proximal promoter CBRs were commercially synthesized (Figure 5C). Corresponding point mutant probes were designed for each by swapping a thymine base, critical for CRX affinity, for an adenine base (ga*T*ta to ga*A*ta). In a 12 μl reaction volume, 35 ng of dsDNA

probes were combined with 300 ng of the human CRX-DBD-MBD fusion protein (0.05 μM final concentration) in 1X PBS for 1 h at 37°C. Reactions were loaded onto an agarose gel and migration of the species was visualized on a 1% agarose gel post stained with GelRed. Data shown are representative of three replicates with the RHO promoter CBR and two replicates with the PDE6B CBR.

3 | RESULTS

3.1 | The CRX activation domain is predicted to be flexible and disordered

The K50 homeodomain transcription factor CRX has been previously shown to orchestrate complex transcriptional regulation in photoreceptor retinal neurons; however, the detailed biochemical mechanism and structural detail of how this critical protein stimulates these effects are not yet apparent. CRX binds to CREs via a conserved N-terminal HD and imparts transcriptional regulation mediated by C-terminal AD, which extends from position 113–284 (Figure 1A).[14,34] Previously, we modeled the CRX DNA binding domain (DBD) structure, suggesting that it formed a compact helical structure similar to other homeodomain-containing transcription factors.[35] Building on this analysis, we assessed the viability of modeling the structure of the transactivation domains of CRX. Unlike the CRX DBD, this portion of CRX shows little homology to known protein structures, and nearly 75% of the protein is predicted to be disordered (Figure 1B). However, we found some disagreement between the prediction methods, especially in the C-terminal AD (Figure 1B). The recent AlphaFold prediction of the proteomes for model species contains entries for human, mouse, rat, and zebrafish CRX, and all of the entries show higher predicted error outside of the DBD.[36,37]

When we plotted the known types of pathogenic mutations from the ClinVar repository at each amino acid position, we observed that the single amino acid substitutions in the activation domain linked to disease are less prevalent than frameshift changes in this region or introduction of a premature stop codon (Figure 1C). In contrast, the single amino acid substitutions are more likely to be pathogenic in the DBD. Other studies of disease-linked amino acid substitutions have noted that frameshift mutations in the activation domain lead to severe disease, but they have not identified significant numbers of point mutants in this region.[38–42] These clinical results are further supported by analysis of the amino acid bias across the different parts of CRX. Within amino acid positions 113–299 of the activation domain are 62 prolines and serines, which amounts to 36% of the amino acids in this region (Figure 1D). The high number of proline residues, which is known to disrupt the secondary structure, and serine, which only moderately favors helical structures, likely leads to the apparent flexibility and disorder in this region.[43] The DBD is rich in basic amino acids, which allows the protein to interact with the backbone and bases of the DNA. Thus, we could not confidently model the structure of any region of CRX outside of the DBD and expand our previous model.

3.2 | Solution shape of the CRX DNA-binding domain

Given the limited structural information on CRX, we attempted to crystallize the homeodomain of CRX, but we have been thus far unsuccessful. Concomitantly, we pursued studies of the structure through small-angle X-ray scattering (SAXS), which is useful for determining the size, shape, and molecular weight of biomolecules.[44] To obtain a homogeneous sample, we fractionated a synthesized peptide of the human CRX homeodomain (CRX-DBD) via size exclusion chromatography. The protein eluted as a single peak and we calculated the molecular weight to be 6.4 kDa with the expected molecular weight from the sequence being 7.4 kDa (Figure 2A). The molecular weight is consistent with this protein being monomeric in solution. We next performed circular dichroism on the peptide to confirm that the peptide was folded. From measurement at 222 nm, we calculated that the CRX-DBD melted at an apparent melting temperature of 56.1 ± 1°C (Figure 2B). After the experiment, we observed precipitation in the cuvette suggesting that the protein aggregated at high temperatures (data not shown). To support the proposal that the CRX-DBD peptide adopts an appropriate fold, we predicted the CD spectrum for several DNA-binding domains with known crystal or NMR structures that

have >60% sequence identity with the CRX homeodomain and the unrelated TFIIIA zinc finger and p53 core domain.[22,24,31,32] The data show that our CD data for the CRX-DBD peptide are more consistent with the structure of a homeodomain (Figure 2C). We then overlaid the predicted CD spectrum of a homology model of the CRX DNA-binding domain over the data and show that the predicted CD spectrum for the CRX DBD is within the error of the data (Figure 2D). Overall, these data suggest that the CRX-DBD peptide is folded in solution and adopts a structure that resembles known and predicted homeodomain structures.

We next submitted the peptide for size exclusion chromatography coupled to SAXS detection. This approach was chosen in part because we observed that the CRX-DBD peptide had stability and solubility issues in equilibrium SAXS experiments (data not shown). From our data collection, we observed that the CRX-DBD peptide migrated as a single peak and we were able to average six frames to get a SAXS data set (Figure 3A,B and Table 1). We calculated an R_g of 11.8 ± 0.2 Å with a molecular weight of 5.7 kDa (98.7% CI 5.0–6.0 kDa), in line with our molecular mass value calculated from an independent size exclusion chromatography experiment. Multi-angle light scattering data were collected on the sample simultaneously with SAXS data and from those data we calculated a molecular weight of
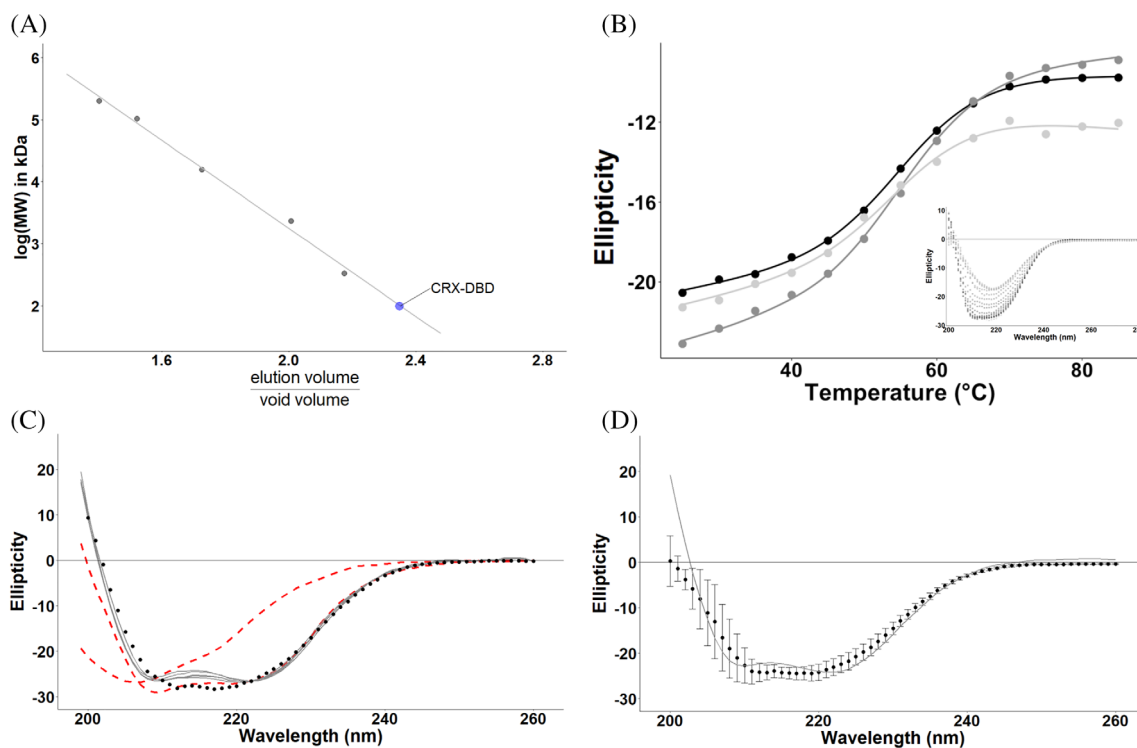

FIGURE 2 Solution properties of the CRX-DBD peptide (A) Elution of CRX-DBD from the size exclusion column. The peptide eluted as a single peak with a calculated molecular weight of 6.3 kDa. These data are consistent with a CRX-DBD monomer. (B) Melting curves of CRX-DBD at 222 nm. Data were globally fit in Calfitter to an equilibrium model and the lines show the fitted data. The calculated apparent melting temperature was 56.1 ± 1 °C. The SSR for the fit was 0.00938. (C) Predicted CD spectra for RCSB ID 2DMS, 3A01, 1FJL, 3CMY, 2MGQ (gray), and 1TF6 and 3EXJ (red lines) overlaid on a representative CD spectrum collected on the CRX-DBD peptide (black points). (D) Predicted CD spectrum for the CRX-DBD homology model (gray line) overlaid on the average CD spectrum for CRX-DBD peptide (black points). Error bars are the SD of the measurements

10.7 ± 1.4 kDa (Figure 3A). The dimensionless Kratky plot for CRX-DBD peptide shows a single peak with a descent toward the X-axis characteristic of a globular protein structure in solution (Figure 3D).

We then predicted SAXS intensity data for three homeodomain structures (1FJL, 3A01, and 3CMY) and fit the data to the SAXS data for CRX-DBD (Figure 4A). For all three structures, the χ^2 was approximately 1, indicating a good match to the data. We then fitted the SAXS data using DAMMIF and DENSS to predict the 3D shape. DAMMIF simulates the overall shape from the SAXS data by

calculating several bead models and then selecting the ones that fit best to further refine and fit the data.[27] DENSS uses iterative structure factor retrieval to calculate the 3D shape and electron density.[26] From these analyses, the envelope resolution calculated in DENSS was 17 Å and the ensemble resolution calculated in DAMMIF was 22 ± 2 Å. The DAMMIF program calculated the mean normalized spatial discrepancy (NSD) of the model to be 0.61 ± 0.04 Å. We fitted these models to the trRosetta homology model to determine the accuracy of the model (Figure 4B, C). The homology model fits the envelope
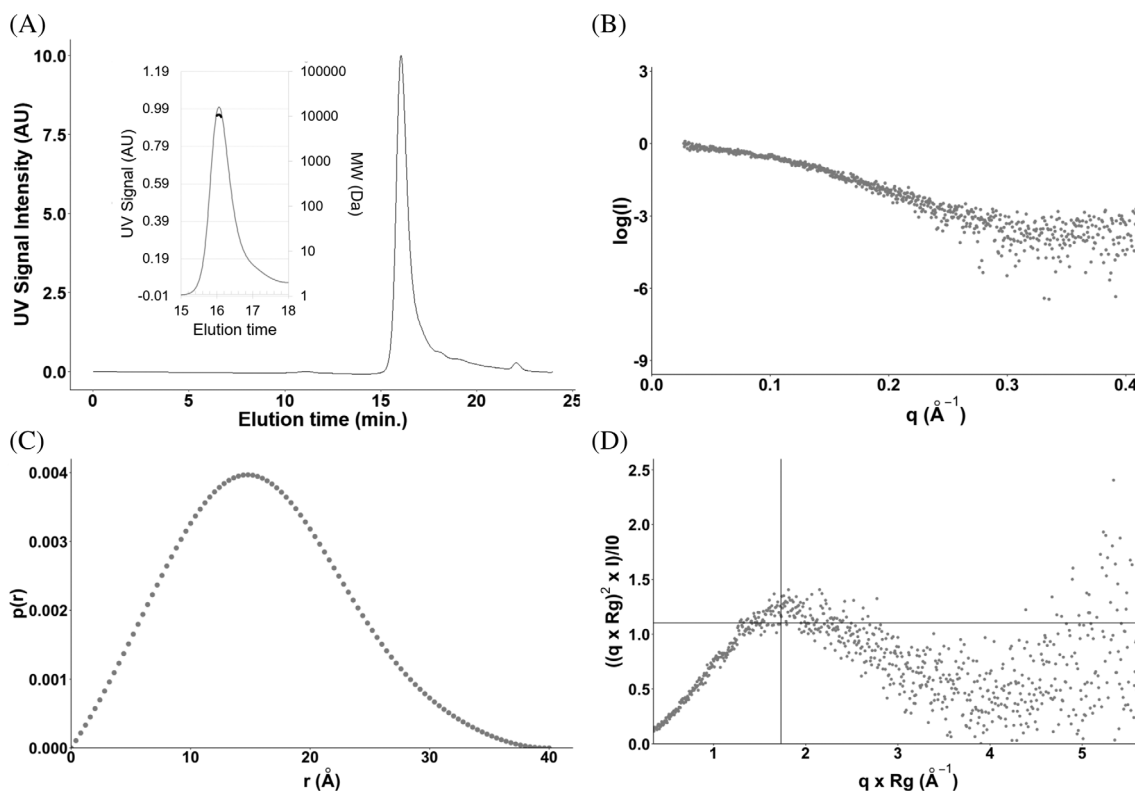
FIGURE 3 SAXS analysis of CRX-DBD (A) UV chromatogram from SEC-SAXS of CRX-DBD peptide. Inset shows the peak and the calculated molecular weight from multi-angle light scattering (points) (B) log of intensity versus momentum transfer for the averaged frames containing CRX-DBD (C) Pair-distance distribution plot (D) Kratky plot of CRX-DBD
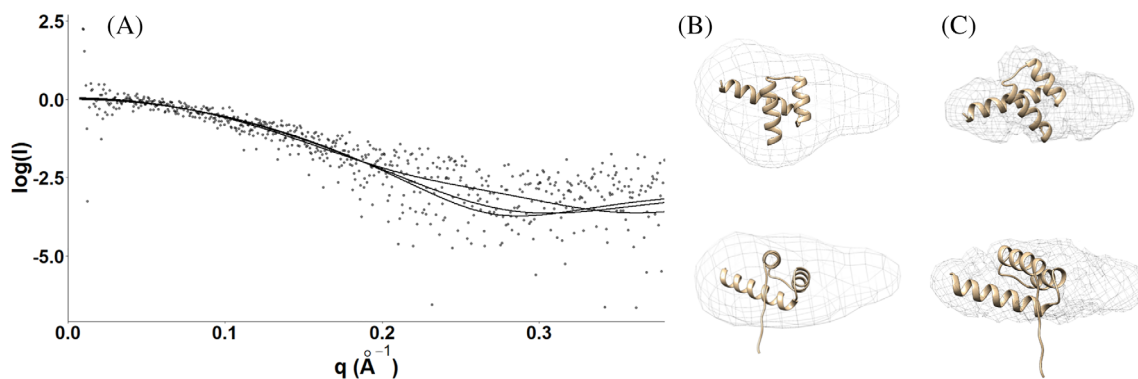
FIGURE 4 Fitting of CRX-DBD SAXS data to model structure (A) Fitting of the FOXS predicted SAXS intensity curve for RCSB ID 1FJL ($\chi^2 = 0.99$), 3A01 ($\chi^2 = 0.99$), and 3CMY ($\chi^2 = 1.02$) to the CRX-DBD intensity data. (B) Envelope of the dummy atom model produced by DAMMIF aligned to the CRX-DBD homology model (C) Electron density map produced from DENSS aligned to the CRX-DBD homology model

from the DAMMIF fitting and the electron density from DENSS. The DBD model contains an N-terminal region that does not have a defined structure. It is possible that this region could adopt several conformations that could occupy the empty portions of the density (Figure 4B and C). These data further support our findings that the CRX-DBD forms a monomer in solution and has a globular structure.

3.3 | The CRX-DBD binds wild-type but not mutated CREs in the human genome

We next analyzed the ability of the CRX-DBD to bind to cis-regulatory elements (CREs) in the human genome using a MBP tagged CRX-DBD fusion protein in mobility shift assays. The MBP tag enhances the shift in the migration of DNA upon binding to the DBD, which aids in resolving bands. Recent ChIP-seq analysis of CRX binding in the genome of human retinal neurons demonstrates CRX-binding regions (CBRs) upstream of the photoreceptor-specific genes *RHO* and *PDE6B* (Figure 5A, B).[10] Further sequence analysis of these regions revealed 5′-GATTA-3′ (5′-TAATC-3′ on reverse strand) putative CRX binding motifs present within each CBR (Figure 5C). The *RHO* upstream region contains two motifs separated by a 4 bp spacer whereas the *PDE6B* upstream region contains a single motif. To determine if these specific motifs within CBRs are functional CREs, we synthesized 35 bp dsDNA oligonucleotides corresponding to the wt sequences as well as point mutant oligos introducing a single base change at a key motif residue (Figure 5C). Incubation of the CRX-DBD-MBD fusion protein with wt *RHO* and *PDE6B* oligos induced a gel shift indicating both regions contained functional CRX CREs (Figure 5D and E). To determine if either of the *RHO* motifs are functional CREs, substitutions of the critical 3rd position of each motif were made. The gel shift was abolished with a base substitution in the first motif, but persisted with a substitution in the second motif, indicating CRE functionality of only one of these motifs (Figure 5D). A similar analysis of the single PDE6B motif demonstrates its functionality as a bona fide CRE (Figure 5E). Collectively, these data demonstrate the functionality of the CRX-DBD protein used in this study. These findings also suggest that CRX binding affinity is dictated by features other than the local base-pair grammar of binding motifs.

4 | DISCUSSION AND CONCLUSIONS

The retina-specific K50 homeodomain transcription factor CRX is required for the differentiation and maturation of photoreceptor neurons in the mammalian retina. While the biological functions of CRX are well described, a detailed exploration of its structural biochemistry is less clear. To address this knowledge gap, we characterized the structure of the CRX DNA-binding domain and its binding to functional DNA CREs. Our data show that the CRX-DBD is monomeric in solution, adopts a compact structure, and is able to bind to specific DNA motifs.

Homeodomain proteins exist in a variety of quaternary structures, including homodimers. In contrast, our size exclusion chromatography

and SAXS data show that the CRX homeodomain is monomeric in solution (Figures 2A and 3A). We do observe two bands in the DNA binding assays, which could be consistent with a DNA-CRX-DBD dimer and a DNA-CRX-DBD ternary complex. Chen et al. in migration shift assays using radiolabeled DNA also observed at high concentrations of their CRX construct that a second band appeared consistent with CRX-dimer formation on the DNA.[45] These data suggest that DNA may mediate dimer formation for CRX or that high concentrations are needed to form the ternary species. However, the cellular relevance and the precise species observed in these previous and the present assays is not clear, indicating the need for further structural and biochemical work.

We further found that a substitution in the third position of the motif abolished binding (Figure 5D and E). However, when two motifs were present, only the first motif was necessary for CRX binding (Figure 5D). These results are consistent with CRX binding to motifs within a broader context than just sequence. Previously, we had shown the effects of DNA methylation on the structure of DNA and the CRX binding motif.[35] In addition, changing the base pair sequence alters the local base stacking energies, which affect the structure and dynamics of the DNA molecule.[46–48] Thus, within the broader context of the DNA molecule in the assay, the second site may not be the preferred binding site for CRX-DBD or the base substitution does not disrupt the contacts between the CRX-DBD and DNA under the assay conditions. Thus, CRX binding motifs may be a marker of potential CRX binding locations but there are additional effects that dictate CRX specificity.

Outside of the CRX-DBD, the activation domain, which is ~60% of the total protein length, is predicted by many approaches to be unstructured (Figure 1B). The lack of clinically relevant single amino acid substitutions in the activation domain further supports this is the case in vivo as small changes in ordered domains are more likely to disrupt the structure (Figure 1C). In the CRX DNA binding domain, 72% of the clinically relevant substitutions are missense substitutions. These substitutions are 65% of the missense variants in the entire protein while the amino acids in the DBD are only 20% of the total length. In the activation domain, 9% of the substitutions are missense while nearly 76% of the pathogenic substitutions in the activation domain are frameshift substitutions (Figure 1C).[38–42] This information supports that the activation domain does not have a specific sequence associated with the structure and requires large changes in sequence to induce effects on protein function. In support of this idea, Chen et al. showed that CRX constructs with single amino acid substitutions in the activation domain showed near-normal trans-activation, while frameshifts and early stops in the activation domain strongly affected activity.[45] These clinical findings support a region with a dynamic structure. A disordered activation domain seems to be at odds with the number of proteins that are known or predicted to interact with CRX in this region.[49] However, flexible binding sites within proteins are recognized with increasing frequency as important to cellular processes, including transcription.[50–52] Given the significance of the activation domain region to gene regulation by CRX, the form and function of the CRX activation domains warrant further study.

FIGURE 5 CRX binds evolutionarily conserved CREs in the human genome. UCSC genome browser visualizations of the human (A) *RHO* and (B) *PDE6B* loci aligned with BAM density plots from an adult human retina CRX ChIP-seq experiment identifying CRX-binding regions (CRBs) in both loci10. Sequences highly conserved between vertebrate species are indicated as green peaks in the PhastCons Conservation data track. (C) Sequences of dsDNA probes used for EMSAs in the *RHO* and *PDE6B* promoter regions. Five base pair predicted CRX motifs within CBRs are indicated in blue lettering. Mutated bases are shown in red and capital letters to indicate the position and identity of the change. EMSA assays for (D) the *RHO* promoter and (E) the *PDE6B* promoter. 34 ng of was used for each assay, and 300 ng purified human MBP-CRX-DBD protein added to lanes indicated with +

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26332.

DATA AVAILABILITY STATEMENT

The SAXS data that support the findings of this study are openly available in SASBDB at https://www.sasbdb.org/ under reference number SASDMA5. Other data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Christopher E. Berndsen 🔟 https://orcid.org/0000-0002-4285-4341

REFERENCES

1. Kolb H. Photoreceptors. In: Kolb H, Fernandez E, Nelson R, eds. *Webvision: the Organization of the Retina and Visual System*. University of Utah Health Sciences Center; 2005.
2. Brzezinski JA, Reh TA. Photoreceptor cell fate specification in vertebrates. *Development*. 2015;142:3263-3273.
3. Bürglin TR, Affolter M. Homeodomain proteins: an update. *Chromosoma*. 2016;125:497-521.
4. de Mendoza A, Sebé-Pedrós A, Šestak MS, et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci U S A*. 2013;110:E4858-E4866.
5. Miller DM, Shen MM, Shamu CE, et al. C. elegans unc-4 gene encodes a homeodomain protein that determines the pattern of synaptic input to specific motor neurons. *Nature*. 1992;355:841-845.
6. White JG, Southgate E, Thomson JN. Mutations in the Caenorhabditis elegans unc-4 gene alter the synaptic input to ventral cord motor neurons. *Nature*. 1992;355:838-841.
7. Schneitz K, Spielmann P, Noll M. Molecular genetics of aristaless, a prd-type homeo box gene involved in the morphogenesis of proximal and distal pattern elements in a subset of appendages in Drosophila. *Genes Dev*. 1993;7:114-129.
8. Koike C, Nishida A, Ueno S, et al. Functional roles of Otx2 transcription factor in postnatal mouse retinal development. *Mol Cell Biol*. 2007;27:8318-8329.
9. Corbo JC, Lawrence KA, Karlstetter M, et al. CRX ChIP-seq reveals the cis-regulatory architecture of mouse photoreceptors. *Genome Res*. 2010;20:1512-1525.
10. Cherry TJ, Yang MG, Harmin DA, et al. Mapping the cis-regulatory architecture of the human retina reveals noncoding genetic variation in disease. *Proc Natl Acad Sci U S A*. 2020;117:9001-9012.
11. Chau KY, Chen S, Zack DJ, Ono SJ. Functional domains of the cone-rod homeobox (CRX) transcription factor. *J Biol Chem*. 2000;275: 37264-37270.
12. Chen S, Wang QL, Nie Z, et al. Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*. 1997;19:1017-1030.
13. Tran NM, Zhang A, Zhang X, Huecker JB, Hennig AK, Chen S. Mechanistically distinct mouse models for CRX-associated retinopathy. *PLoS Genet*. 2014;10:e1004111.
14. Tran NM, Chen S. Mechanisms of blindness: animal models provide insight into distinct CRX-associated retinopathies. *Dev Dyn*. 2014; 243:1153-1166.
15. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980-D985.
16. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*. 2007;35:W460-W464.
17. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res*. 2018;46:W329-W337.
18. Dass R, Mulder FAA, Nielsen JT. ODiNPred: comprehensive prediction of protein order and disorder. *Sci Rep*. 2020;10:14780.
19. Gasteiger E et al. Protein identification and analysis tools on the ExPASy server. In: Walker JM, ed. *The Proteomics Protocols Handbook*. Humana Press; 2005:571-607.
20. Hopkins JB, Gillilan RE, Skou S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J Appl Cryst*. 2017;50:1545-1553.
21. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer; 2010.
22. Wilson DS, Guenther B, Desplan C, Kuriyan J. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell*. 1995;82:709-719.
23. Miyazono K-I, Zhi Y, Takamura Y, et al. Cooperative DNA-binding and sequence-recognition mechanism of aristaless and clawless. *EMBO J*. 2010;29:1613-1623.
24. Birrane G, Soni A, Ladias JAA. Structural basis for DNA recognition by the human PAX3 homeodomain. *Biochemistry*. 2009;48:1148-1155.
25. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. FoXS, FoXSDock and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res*. 2016;44:W424-W429.
26. Grant TD. Ab initio electron density determination directly from solution scattering data. *Nat Methods*. 2018;15:191-193.
27. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J Appl Cryst*. 2009;42: 342-346.
28. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25:1605-1612.
29. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A*. 2020;117:1496-1503.
30. Mazurenko S, Stourac J, Kunka A, et al. CalFitter: a web server for analysis of protein thermal denaturation data. *Nucleic Acids Res*. 2018; 46:W344-W349.
31. Nolte RT, Conlin RM, Harrison SC, Brown RS. Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proc Natl Acad Sci U S A*. 1998;95:2938-2943.
32. Malecka KA, Ho WC, Marmorstein R. Crystal structure of a p53 core tetramer bound to DNA. *Oncogene*. 2009;28:325-333.
33. Drew ED, Janes RW. PDBMD2CD: providing predicted protein circular dichroism spectra from multiple molecular dynamics-generated protein structures. *Nucleic Acids Res*. 2020;48:W17-W24.
34. Ruzycki PA, Tran NM, Kefalov VJ, Kolesnikov AV, Chen S. Graded gene expression changes determine phenotype severity in mouse models of CRX-associated retinopathies. *Genome Biol*. 2015;16:171.

35. Hossain RA, Dunham NR, Enke RA, Berndsen CE. In silico modeling of epigenetic-induced changes in photoreceptor cis-regulatory elements. *Mol Vis*. 2018;24:218-230.

36. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589. doi:10.1038/s41586-021-03819-2

37. Tunyasuvunakool K, Adler J, Wu Z, et al. Highly accurate protein structure prediction for the human proteome. *Nature*. 2021;596:590-596. doi:10.1038/s41586-021-03828-1

38. Swain PK, Chen S, Wang QL, et al. Mutations in the cone-rod homeobox gene are associated with the cone-rod dystrophy photoreceptor degeneration. *Neuron*. 1997;19:1329-1336.

39. Freund CL, Gregory-Evans CY, Furukawa T, et al. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell*. 1997;91:543-553.

40. Freund CL, Wang QL, Chen S, et al. De novo mutations in the CRX homeobox gene associated with Leber congenital amaurosis. *Nat Genet*. 1998;18:311-312.

41. Sohocki MM, Sullivan LS, Mintz-Hittner HA, et al. A range of clinical phenotypes associated with mutations in CRX, a photoreceptor transcription-factor gene. *Am J Hum Genet*. 1998;63:1307-1315.

42. Sohocki MM, Daiger SP, Bowne SJ, et al. Prevalence of mutations causing retinitis pigmentosa and other inherited retinopathies. *Hum Mutat*. 2001;17:42-51.

43. Pace CN, Scholtz JM. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys J*. 1998;75:422-427.

44. Putnam CD, Hammel M, Hura GL, Tainer J. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys*. 2007;40:191-285.

45. Chen S, Wang QL, Xu S, et al. Functional analysis of cone–rod homeobox (CRX) mutations associated with retinal dystrophy. *Hum Mol Genet*. 2002;11:873-884.

46. Protozanova E, Yakovchuk P, Frank-Kamenetskii MD. Stacked-unstacked equilibrium at the nick site of DNA. *J Mol Biol*. 2004;342:775-785.

47. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res*. 2006;34:564-574.

48. Li G, Quan Y, Wang X, et al. Trinucleotide Base pair stacking free energy for understanding TF-DNA recognition and the functions of SNPs. *Front Chem*. 2018;6:666.

49. Peng G-H, Chen S. Crx activates opsin transcription by recruiting HAT-containing co-activators and promoting histone acetylation. *Hum Mol Genet*. 2007;16:2433-2452.

50. Fuxreiter M, Tompa P, Simon I, Uversky VN, Hansen JC, Asturias FJ. Malleable machines take shape in eukaryotic transcriptional regulation. *Nat Chem Biol*. 2008;4:728-737.

51. Staby L, O'Shea C, Willemoës M, Theisen F, Kragelund BB, Skriver K. Eukaryotic transcription factors: paradigms of protein intrinsic disorder. *Biochem J*. 2017;474:2509-2532.

52. Brodsky S, Jana T, Mittelman K, et al. Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Mol Cell*. 2020;79:459-471.e4.