# Modelling protein-DNA interactions: Bend and Twist until it fits

**Marc van Dijk**

ISBN/EAN 978-90-393-5280-9

Doctoral thesis
Modelling protein-DNA interactions: Bend and Twist until it fits
Marc van Dijk
NMR Spectroscopy Research Group, Bijvoet Center for Biomolecular Research,
Utrecht University, The Netherlands
February 2010

Financial support for the publication of this thesis was kindly provided by the
Netherlands Bioinformatics Centre (NBIC)



Printed in the Netherlands by Ridderprint Offsetdrukkerij B.V.

# Modelling protein-DNA interactions: Bend and Twist until it fits

## Modelleren van Eiwit-DNA interacties: buigen en draaien totdat het past

(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. J.C. Stoof, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 8 februari 2010 des middags te 12.45 uur

door

**Marc van Dijk**

geboren op 5 december 1977
te Gouda

Promotoren:

Prof. dr. A.M.J.J. Bonvin
Prof. dr. R. Boelens

Beoordelingscommisie:

Prof. dr. R. Kaptein
Prof. dr. G.T. Barkema
Prof. dr. G.W. Vuister
Prof. dr. M. Zacharias
Dr. P.A. Bates

**“The task is not so much to see what no one has yet seen; but to think what nobody has yet thought, about that which everybody sees.”**

*Erwin Schrödinger (1887-1961)*

*Opgedragen aan:*
Mijn vader, mijn moeder
Cécile mijn steun en toeverlaat

# Table of Contents



1) Full colour figures are marked with an asterisk (*) throughout the text

# List of abbreviations

| | |
|---|---|
| 3D-DART | 3DNA-driven dna analysis and rebuilding tool |
| AANT | amino acid-nucleotide interaction database |
| AIR | ambiguous interaction restraint |
| CAPRI | critical assessment of prediction of interactions |
| CNS | crystallography and NMR system |
| DNA | deoxyribonucleic acid |
| HADDOCK | high ambiguity driven docking |
| HSSP | homology-derived secondary structure of proteins |
| JUMNA | junction minimisation of nucleic acids |
| MD | molecular dynamics |
| NAB | nucleic acid builder |
| NAMOT | nucleic acid modelling tool |
| NMA | normal mode analysis |
| NMR | nuclear magnetic resonance |
| NOE | nuclear Overhauser effect |
| PCA | principle component analysis |
| PDB | protein data bank |
| PEDANT | protein extraction, description and analysis tool |
| PFAM | protein families database |
| PROSITE | database of protein domains, families and functional sites |
| PSVS | protein structure validation software suite |
| RCSB | research collaboratory for structural bioinformatics |
| RDC | residual dipolar coupling |
| RMSD | root-mean-square-deviation |
| RNA | ribonucleic acid |
| SCOP | structural classification of proteins |
| SMART | simple modular architecture research tool |

# Introduction

Chapter 1

## The blueprint of life

We celebrated in 2009 the 200th birthday of Sir Charles Robert Darwin (12 February 1809) and the 150th anniversary of his publication "On the Origin of Species" (68). In his extensive survey of the geographical diversity in living organisms aboard the HMS Beagle, Darwin described by observation, the many ways that species seem to have adapted to their environments. Many species appear to have a common ancestor and only differ from each other depending on the environment they live in. In his publication, Darwin postulated that under the influence of a changing environment only those species with specific traits that enable them to survive under the new conditions would thrive while others perish. Somehow these favourable traits are preserved and enriched in the population, which eventually, can lead to new species. His work provided a revolutionary view on the origin of the living world. The evolution of life from common ancestors over time driven by the process of natural selection is the corner stone of the modern evolutionary theory.

Not long after Darwin's publication, the Augustinian priest Gregor Johan Mendel published his experimental work on inheritance of traits in pea plants (217). Mendel performed extensive experiments on the distribution of traits over many generations of pea plants and derived the underlying rules of inheritance. Unaware of Mendel's work, the Dutch botanist Hugo de Vries performed similar inheritance experiments. He introduced the terms "genes" and "mutations" (335). Although Mendel's work initially received little attention he is now commonly regarded as the father of modern genetics.

Although the principles of inheritance became apparent, the mechanism by which the trait information is passed from generation to generation was still unknown. It was the Swiss physician Friedrich Miescher who first isolated DNA from discarded surgical bandages in 1869. The first insights into the molecular composition of DNA came from Phoebus Levene in 1919: he discovered the nucleotide unit and its components the base, sugar and phosphate backbone, and proposed that they are linked together as a chain using the phosphate groups (186). The role of DNA as the carrier of genetic information was discovered later on in a series of transformation and reproduction experiments in bacteria and phage's (16,128,192). With the DNA recognized as the carrier of genetic information, the focus then shifted to the elucidation of its structure.

## The hunt for genes and the DNA structure

James D. Watson and Francis Crick proposed the now well accepted DNA double-helix model in 1953 based on a fiber diffraction image obtained by Rosalind Franklin, Maurice Wilkins and co-workers in May 1952 (341). The double helix appears intriguingly simple and elegant in design: A long linear polymer composed out Levene's repeating nucleotides. The nucleotides only differ from each other in their base unit of which there are four different ones arranged in pairs as proposed by Chargaff (52-54): Adenine (A) pairs with thymine (T) and guanine (G) pairs with cytosine (C). The four bases, when grouped in codons of three, give 64 possibilities, more than enough to uniquely identify the twenty standard amino acids (62). It is with this reasoning that Francis Crick laid out the "Central Dogma" of molecular biology (63): the genetic information encoding a protein is grouped together as a number of codons on the DNA. This sequence is transcribed into a messenger RNA that is subsequently translated by the ribosome into a chain of amino acids. The discovery of the DNA double helix and the foundation of the "Central Dogma" sparked a new wave in science. The research focus again shifted,

this time to the identification of genes in various genomes and to the functional annotation of the gene products.

The DNA of the bacteriophage φX174 (271) was the first to be sequenced using the technique developed by Frederick Sanger (272). In the 25 years that followed, numerous genomes have been sequenced (26), culminating in the first draft of the human genome in 2001 (172,329) as the cumulative result of the combined efforts of many laboratories worldwide. The vast amount of sequence data that resulted from these efforts revealed the location of known proteins on the genetic material, but also dramatically increased the number of putative proteins. With the "Central Dogma" in mind the question arises what are the functional protein products of all those genes and how is the process from gene to protein is regulated? The functional annotation of genes is greatly facilitated by biochemical experiments, notably mutagenesis experiments. However, the true appreciation of a protein's function comes from the study of its molecular structure and interactions with other biomolecules as it is with these interactions that proteins fulfil their function.

The two classical structure determination methods, X-ray crystallography and Nuclear Magnetic Resonance spectroscopy (NMR) are the main experimental methods used to determine the structure of proteins and protein complexes (38,69). X-ray scattering enabled Watson and Crick to propose the double-helix DNA. The first single-crystal structure of a DNA molecule was solved in 1979 by Wang and co-workers (338). X-ray diffraction studies on protein-DNA complexes date back as far as 1974 with the structural studies on the *Escherichia coli Lac* repressor (299) followed by the first DNA binding proteins (12,241,342). The first high-resolution structures of protein-DNA complexes were obtained for the *Eco*RI endonuclease (104), bacteriophage 434 repressor (2,11), λcI repressor (144), DNAse I (301), *E. coli trp* repressor (240) and the Klenow fragment of *E. coli* DNA polymerase I (105).

NMR spectroscopy provides a high-resolution view on the structure and dynamics of proteins and protein-DNA complexes and is readily used to map interaction interfaces. The first solution structures of DNA binding proteins were obtained for the *E. coli lac* repressor (150), *Drosophila Antennapedia* homeodomain (255) and several zinc 'finger' domains (181,248). The first structural studies on protein-DNA interactions using NMR were also performed on the *E. coli lac* repressor (36,60,148,297).

As the number of solved structures of DNA and protein-DNA complexes increased, the DNA manifested itself as a dynamic and complex molecule. The specific recognition of DNA by proteins appears to result from a subtle interplay between the DNA sequence and the conformational changes resulting from complex formation.

The development of more powerful high-throughput molecular biology techniques revealed the interplay between proteins and DNA. This provided a new view on the processes occurring in a living cell showing the way biomolecules interact with one another as a response to a changing environment. Much like Darwin's view of the world, atomic structures were no longer viewed as isolated entities but as parts of large dynamic networks.

## Beyond structure: dynamics and interactions

As the efforts to solve the atomic structures of components of interaction networks increased, the limitation of the experimental methods became apparent. Both X-ray crystallography and NMR spectroscopy can be quite labor intensive, which imposes constraints considering the large number of possible complexes that greatly exceeds the number single proteins solved (153).

Obtaining quality crystals for X-ray experiments or a pure protein sample of sufficient concentration and solubility for NMR can be a daunting task. The current size limit for NMR (50-70 kDa) does not allow for routine studies of large complexes. Next to these limitations there is a large group of complexes that is extremely difficult to solve using both techniques. Among these are transient, short-lived complexes, membrane associated complexes and protein nucleic-acid complexes, all of which are biologically very interesting.

DNA-interacting proteins fulfil an important role in the biomolecular interaction networks of the living cell. If changes in the environment of the cell require adaption of its protein content, it is eventually up to the DNA-interacting proteins to induce changes in the level of transcription. Furthermore, sophisticated protein machineries guard the DNA against hazardous influences that may damage it. A disruption of this delicate balance of regulation can lead to severe consequences for the cell, ranging from the inability to perform certain functions to cancer or cell death. An in-depth understanding of the mechanism underlying these regulation processes is only possible by studying them at atomic detail. Although the number of solved protein-DNA complexes in the RCSB protein database (30) is steadily growing, the putative number of DNA-binding proteins and protein interaction motifs on the DNA (207) is still much larger. Examination of genes that are functionally assigned in the PEDANT database (106) shows that typically 2-3% of a prokaryotic genome and 6-7% of a eukaryotic genome encodes DNA-binding proteins.

Considering the limitations of experimental techniques in solving the 3D structures of biomolecular complexes, it becomes difficult or impossible to study many of the complexes in interaction networks. Here, complementary computational techniques, notably docking have proven to be a valuable asset (133). Docking is the art of modelling the complex in its "bound" state from its "unbound", free components, using a variety of computational algorithms. Docking, therefore, aids in the verification of hypotheses and provides a fast method to plan future experiments based on initial models or screening of targets. Protein-protein and protein-ligand docking (e.g. pharmaceuticals) is nowadays readily used in both academia and industry. The development of protein-DNA docking methods, however, lags behind.

## Outline of this thesis

This short introduction should have clearly illustrated the importance of protein-DNA interactions. Detailed knowledge of the molecular structure of these complexes provides valuable insights into their function and impacts on many different fields such as medical sciences, molecular biology and pharmacology. The focus of this thesis is on the development of an effective protein-DNA docking method by extending the capabilities of the docking program HADDOCK (86) developed in our group. HADDOCK is a data-driven docking method that allows for explicit flexibility. The first in particular means that HADDOCK is able to directly use experimental information about a complex for the benefit of docking. This ensures that possible solutions are in agreement with experiments. The latter provides the method with the potential to deal with (modest) conformational changes upon complex formation.

**Chapter 2** discusses the characteristics of protein-DNA systems from a docking perspective, provides a historical overview of the protein-DNA docking field and defines the main challenges that dominate the field. The chapter concludes with an overview of HADDOCK and how its special features can be used in protein-DNA docking. Based on this, a protocol for protein-DNA docking

using HADDOCK is defined. In **Chapter 3**, this docking protocol is put to the test on a set of three dimeric transcription factor-DNA complexes in their monomeric form. By including a variety of experimental and/or bioinformatics data, together with a flexible description of the DNA, the method outperforms rigid-body docking and reproduces many of the interface contacts and the specific conformation of the DNA. **Chapters 4** and **5** set the stage for an expansion of the method to a larger variety of protein-DNA systems. **Chapter 4** describes 3D-DART, a method for the generation of custom DNA models by allowing full control over the bend angle between successive base pairs. The method is made available through a convenient web interface. **Chapter 5** describes a dedicated non-redundant protein-DNA docking benchmark containing 47 test cases covering a variety of challenging systems. This benchmark should be a useful tool for method development and validation. **Chapter 6** combines all tools and protocols described in **chapters 2** to **5** into a large-scale docking effort defining the limits of what is currently achievable in the field of protein-DNA docking. **Chapter 7** concludes this thesis by giving a perspective on the methods and tools presented and the protein-DNA docking field in general.

# Docking protein-DNA complexes

*Based in part on:*

Protein-DNA docking:
The tricks of an emerging trade

Marc van Dijk
Alexandre M.J.J. Bonvin

*Manuscript submitted for publication*

Chapter 2

How do biomolecules interact to perform their function? is one of the central questions in system biology to date. The study of biomolecular complexes at atomic resolution by NMR and X-ray crystallography are vital in answering this question. In the last decade the tool chest of the structural biologist has been extended with a set of powerful computational techniques notably docking. Docking allows for the study of complexes that would otherwise be difficult to impossible to solve using experimental techniques. Docking of protein-protein complexes is nowadays widely used. In contrast, the field of protein-DNA docking has seen little development. However, with a renewed interest in protein-DNA complexes new docking methods are put forward and proven protein-protein docking concepts are extended to deal with these systems. In this chapter we look at the docking field from the perspective of protein-DNA systems. With a discussion of the features of protein-DNA complexes it becomes clear that conformational changes and the identification of the correct interaction interface on the regular DNA helix, are the main challenges dominating the field. A survey of the early methodology shows that the initial focus was, mostly, on the identification of the correct interfaces using small, well-defined test systems, often using rigid-body docking. Recent promising developments, however, focus more and more on the proper treatment of conformational changes upon complex formation in both protein and DNA. The chapter concludes with a discussion of the HADDOCK data-driven docking approach and the way that it could potentially deal with the main challenges that dominate the protein-DNA docking field.

## Introduction

High-throughput analysis is commonly applied in many areas of life science such as genome sequencing, micro-array technologies, protein expression and purification techniques and computational methods. The large-scale sequence analysis of genomes has become a routine exercise that has enriched the scientific knowledge with a vast amount of sequence data (13). These data have revealed the location of know proteins on the genetic material but also dramatically increased the number of putative proteins. The total protein content of a living cell, also known as its proteome, is dynamic in nature as proteins interact with one another and other molecules, and their number and function vary in response to changes in their environment. As a consequence, a lot of attention is focused on system biology that studies biomolecules and the way they work together to fulfil their tasks.

The first stage in such studies involves determining which interactions occur in the pool of biomolecules using various biochemical and biophysical techniques (51). These, however, do not provide any insight into the nature of the interactions at atomic resolution, the second stage in a system biology study. This is still, mostly, the domain of biophysical methods and mainly X-ray crystallography and Nucleic Magnetic Resonance (NMR) spectroscopy. Although these methods will remain indispensable to the structural biologist they also have their limitations. Both methods are quite labour intensive, which imposes constraints considering the large number of possible complexes that greatly exceeds the number single proteins solved (153). Obtaining quality crystals for X-ray experiments or a pure protein sample of sufficient concentration and solubility for NMR can be a daunting task. The current size limit for NMR (50-70 kDa) does not allow for routine studies of large complexes. Because of these limitations there is a large group of complexes that are extremely difficult to solve using both techniques. Among these are transient, short-lived complexes, membrane associated complexes and protein nucleic-acid complexes, all of which are biologically very interesting.

Computational docking provides an appealing alternative for those cases where experimental techniques are unsuccessful in solving the 3D structure of a biomolecular complex. Docking is the art of modelling the complex in its "bound" state from its "unbound", free components, using a variety of computational algorithms. Computational docking has become a valuable addition to the structural biologists tool chest. It allows for the study of complexes that would otherwise be difficult or impossible to solve and provides a fast method to generate new hypotheses and plan future experiments based on initial models or by screening for targets. Protein-protein docking is widely used in academia and protein-small ligand docking has become common practice in industry for the purpose of drug discovery (reviewed in (111,120,146,167,184,296))

The modelling of protein-DNA systems, however, lags behind. The first, pioneering, work in this field dates back about 20 years with the modelling of the phage 434 Cro and Catabolic gene Activator Protein (CAP) DNA systems using pre-bent DNA and an interactive molecular graphics approach in which electrostatic complementarity was considered (212,340,343). One of the first true cases of protein-DNA docking was performed by Kasinos *et al* (151). They reported a graph theory-based method, which was successfully applied to model the Met repressor-DNA complex. All of these early attempts, however, relied heavily on accurate knowledge of the interfaces, including the contacts made, which are typically difficult to obtain experimentally. After these initial attempts the field received only little attention. This, however, is

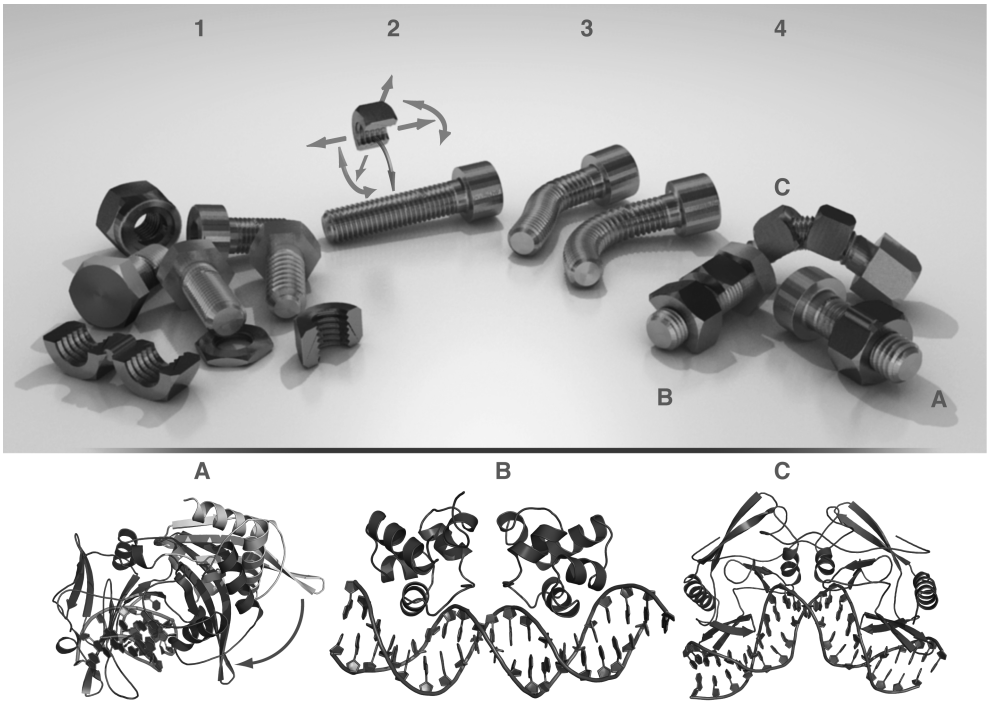



***Figure 2.1.** Schematic representation of the main challenges in protein-DNA docking; **1**) Selecting starting structures for the protein (red bolts) and the DNA (grey screws), which can exist in multiple conformations. **2**) Searching conformational space aimed at assembling the interaction interfaces by rigid body translations and rotations (red arrows). **3**) Dealing with conformational change during the conformational search (DNA, deformed screws). **4**) Ranking the solutions and selecting the "correct" solutions among the many different poses. Examples of complexes: **A**) restriction endonuclease MVAI (20aa) that undergoes a hinge motion upon DNA interaction from the open state (yellow) to the closed state (red); **B**) Dimeric bacteriophage 434 Cro transcription factor (3cro); and **C**) I-PpoI homing endonuclease (1a74), heavily kinks DNA upon complex formation. The schematic figure in the top panel was generated using Blender (www.blender.org) and the figures in the bottom panel were generated using Pymol (DeLano Scientific, www.pymol.org).



changing as the vital role of protein-DNA interactions in regulating gene expression and guarding genome integrity has become imminent (90). New protein-DNA docking methods are pioneered and many methods that were originally developed for protein-protein and protein-ligand docking are now modified to deal with protein-nucleic acid systems. Although the number of solved protein-DNA complexes in the RCSB protein database (30) is steadily growing, the putative number of DNA-binding proteins and protein interaction motifs on the DNA (207) is still much larger. Examination of genes that are functionally assigned in PEDANT (106) shows that typically 2-3% of a prokaryotic genome and 6-7% of a eukaryotic genome encodes DNA-binding proteins. It is clear that there still is a large body of unknown DNA binding proteins. They often fulfil a vital role in the living cell and can be difficult to solve using experimental techniques. Considering this, it is expected that efficient protein-DNA docking methods will become of vital importance.

In this chapter we focus on the protein-DNA docking methodology. Although many principles also apply to other protein-nucleic acid systems there are still considerable differences and we will not venture into these here. We start with a brief description of the characteristics of protein-DNA systems that distinguishes them from protein-protein complexes. We then examine the four main problems associated with the development of a successful docking approach from a protein-DNA perspective and address the way in which current methods try to deal with them (Figure 2.1). These include:

1. The selection of molecules, and their conformation, to start the docking process.
2. The search through conformational space aimed at assembling the correct interface(s).
3. The ways to deal with conformational changes during complex formation.
4. The ranking of solutions and selection of the relevant ones.

We finish this chapter with a discussion of the HADDOCK data-driven docking approach, describing the unique features of this method and the way it could potentially be used to effectively deal with the four main challenges mentioned above. Based on this we formulate a protein-DNA docking protocol for HADDOCK.

## Protein-DNA complexes, general implications

Much of the current docking methodology has been developed for the docking of protein-small ligand and protein-protein systems. Although there are many parallels to protein-DNA systems there are also distinct differences that affect the docking methodology used. In this section we outline the main characteristics of protein-DNA systems and introduce general concepts that will be used in the description of the methodology later on.

A protein-DNA system is heterogeneous in nature: It contains DNA as a regular, helical polymer with a highly negatively charged sugar-phosphate backbone and a protein with a DNA binding interface predominantly composed of above average polar and positively charged amino acids (R, T, S, K (142,183)). Although protein and DNA are quite distinct from each other both chemically and structurally, the chemical and shape complementarity of their interface is central to their interaction (122,260). As the list of protein–DNA complexes with solved three-dimensional (3D) structures continues to expand, the factors underlying their interaction become better and better characterized (reviewed in (243,249)).

### Protein-DNA interfaces

A systematic study of the protein-DNA complexes in the RCSB protein databank resulted in several classification schemes (197,253,275,288) based on the topology of the complex, the protein structure and function and the DNA. From these studies it became clear that DNA binding proteins interact with DNA using all common secondary structure elements (α-helix, β-sheet, turn) and a variety of loop structures, alone or in combination. These secondary structure elements, when combined in DNA recognition domains on the interaction surface of the protein, often align with the DNA grooves in a distinct topology, following the direction of the helix (352).

Protein-DNA interfaces are more polar in nature and contain more hydrogen bonds than protein–protein interfaces (141,142,198,206,228,285,349). Indeed, hydrogen bonding is the major element in sequence specific recognition, since both bases and amino acids have hydrogen donor/acceptor potentials. In most complexes the majority of the hydrogen bonds are non-specific, stabilizing contacts, involving the

phosphodiester backbone: (243) only a small number is targeted to the bases (183). Hydrophobic interactions (27) and water-mediated hydrogen bonds also play a very important role in DNA binding.

There are up to twice as many bridging waters in protein-DNA than in protein-protein complexes (142). They appear, however, to play a more important role in the stabilisation of the protein-DNA complex rather than in specific recognition (198). The ordered spine of water molecules located in the minor groove of B-DNA and the major groove of A-DNA is well recognized (141,258). The B-DNA major groove is too wide to retain the same well-ordered water network and, as a consequence, water molecules are found interacting alone or in pairs with the nucleotide bases. This ordered pattern of water molecules around the DNA is so conserved that their location can be quite accurately predicted ((277) http://www-ibmc.u-strasbg.fr/arn/sws.html). The spine of hydration seems to be a common feature of A-T-rich regions and is presumed to stabilize the DNA conformation (258). Changes in base sequence and base morphology result in different hydration patterns (277,278,280). These have been proposed to be involved in the initial screening by a protein to find a favourable interaction site (107,349), a form of indirect recognition. Furthermore, water molecules at protein-DNA interfaces have been proposed to buffer the electrostatic repulsions between phosphate groups of the DNA and electronegative groups of the protein (258), resulting in a screening of unfavourable electrostatic contacts.



**DNA recognition and specificity**

When considering specificity, the major groove can form more specific interactions than the minor groove. This is due to the easily accessible bases in the major groove and to the fact that hydrogen bond donor and acceptor patterns are unique for each base in the major groove but not in the minor groove. It is for instance not possible to distinguish AT from TA or GC from CG in the minor groove. This is also the reason why B-DNA is a better candidate for specific recognition than A-DNA and why most transcription factors are major groove binders (183,198,204,205,228,246). Although a lot of information about protein-DNA complex formation has become available there is still no clear alphabet of specific amino acid to base interactions. These interactions seem to be widely distributed in space, i.e. the same pair may interact using a variety of geometries. The protein-DNA "recognition code" seems to be degenerate in both directions (25,59); in other words, each DNA base could be recognised by a limited set of amino acids and vice-versa. This two way degeneracy has led some groups to postulate that there exists no recognition code at all (213). Nevertheless, a recognition code that is probalistic in both directions, "P-code" (77,198,205,229), has been proposed from clear base-amino acid preferences (Arg-G, Asn-A, Lys-G (6,158,204,286)). These preferences have been shown to be useful in docking as further discussed later on. The DNA recognition mechanism is further complicated by the effect of the DNA intrinsic flexibility: the latter does play a role in the establishment of the initial encounter complex and, therefore, is responsible for part of the specificity/recognition (57,115,132,162,163,180,282-284). This is also known as the indirect readout (88,240). If we consider that a transcription factor, for instance, can interact with many slightly different versions of the same operator this mechanism becomes understandable (250,260,304,306).

**Conformational changes**

In spite of the various motifs used by DNA-binding proteins, the protein and DNA surfaces are always able to closely match each other (142). This 'snug fit' between

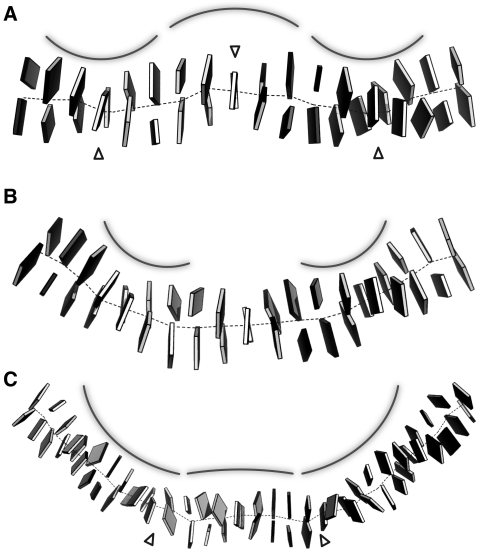



**Figure 2.2.** Three types of DNA conformational adaptation in protein-DNA complexes schematically represented as a two block per base pair Calladine-Drew plot. The black dotted line defines the main helical path and the open triangles the location of local strong bending (kinking). The interacting protein surfaces are represented as shaded black curves. (**A**) Type 1: multiple local changes that cancel each other, effectively leaving the DNA helical path straight (glucocorticoid receptor, 1r40) (**B**) type 2: cumulative small local changes that, together, effectively bend the DNA in a smooth way (MATa1/MATα2 homeodomain, 1yrn) and (**C**) type 3: severe local distortions, like kinks, that effectively bend the structure (catabolite activator protein (CAP), 103t). The figures were generated using 3DNA (194,195) and Pymol (DeLano Scientific, www.pymol.org).

protein and DNA is, in part, facilitated by the ability of the DNA to undergo conformational changes that appear to be coded in specific nucleotide sequences (260). Indeed, subsequent to the initial binding of the protein, the DNA can undergo a variety of conformational changes like groove compression or expansion, unwinding of base pairs or DNA bending (7,82,142,183,228,246,274,312,313,353). The extend of DNA conformational changes depends on its sequence (82,205,210,287,303) and on the number, type, size and interface patch-count (35) of the interacting protein(s).

In contrast to a protein, which is stabilized by multiple long-range stabilizing interactions between side-chains, the DNA lacks such long-range stabilizing interactions. The conformation of a given DNA sequence is a cumulative result of local base stacking events. The base type and sugar-phosphate backbone conformation of an individual base pair and its nearest neighbours as well as the local environment (hydration, ligand interaction) determine the stacking behaviour. As a result, DNA exhibits a flexible behaviour in which it is able to easily change its local conformation in response to changes in the environment. This conformational adaptation is clearly visible in protein-DNA complexes and has been categorized into three main groups (82):

1. Multiple local changes that cancel each other, effectively leaving the DNA helical path straight (Fig. 2.2A).
2. Cumulative small local changes that, together, effectively bend the DNA (smooth bending, Fig. 2.2B).
3. Severe local distortions, like kinks, that effectively bend the structure (Fig. 2.2C).

The extend of conformational changes is in part intrinsic to the DNA and in part protein-induced (83). The DNA structure can be deformed quite dramatically before the internal energy rises unfavourable, a feature proposed to be used by proteins to deform DNA in an economical way (160). Often, the DNA structure has to be deformed by the protein in order for it to gain easier access to the sugar-phosphate backbone or specific bases like in the case of a restriction enzyme or, a transcription factor interacting with several operator sites on the DNA. An example of the former case is the homing endonuclease I-PpoI (1a74, Fig. 2.1C) that kinks the DNA for easier cleavage across the

minor groove. An example of the latter case is the bacteriophage 434 Cro transcription factor (3cro, Fig. 2.1B) where a central DNA spacer is over wound and bent to bring the two adjunct operator sites within range of the DNA binding motifs.

The mechanism of DNA conformational changes has been the topic of a wide range of studies (reviewed in (64,94,254,315)). Consistent relations between base types and flexibility are emerging, such as, for example, A-tracts and pyrimidine-purine base-steps that have a high intrinsic flexibility, more easily allowing for bending and kinking (82). Clear relationships between base pair step conformations expressed in terms of Roll, Tilt, Twist, Slide and Shift parameters (Figure 2.3B) and DNA conformational changes have been identified in free DNA and protein-DNA complexes (22,80-82,116,215,239,305,307,344,358).

Next to the DNA, also the protein often changes its conformation when interacting with its DNA target. The mechanism it utilises in this process can be described by various models: the "key and lock" model, the "induced-fit" model, the "conformational equilibrium" model and the "dynamic shift" model (43,242). The first and second are classic descriptions. The third implies that the protein behaves as an ensemble of local and global conformations sampling the (pseudo)-bound conformation. The dynamic shift model proposes that DNA binding causes a change in the probability distribution of the ensemble of native states. An analysis of different protein-DNA complexes for which more than one bound and free form exist showed that all above described models are actually found (118). Many of these models require that a conformational change takes place in the protein in going from an unbound to a bound conformation. These changes involve subtle side-chain and or backbone rearrangements in or near the interface, loop rearrangements, domain rearrangements and even disordered to order transitions (124,142,175,228). More than in protein-protein systems, DNA binding proteins often occur in a partly disordered state when in their unbound conformation; this is possibly caused by a small hydrophobic core and uncompensated buried charges, a likely consequence of a large binding interface and of the necessity to bind to a negatively charged polymer backbone (91,319,350). This type of disorder to order transitions have been proposed as a mechanism that allows a protein to adapt to multiple targets. This, again, makes sense in the case of, for instance, transcription factors that often bind to a population of related operator sequences.

**Implications for protein-DNA docking**

The above-mentioned features of protein-DNA complexes impose considerable challenges to protein-DNA docking methods, notably:

1. The omnipresence of conformational changes in both protein and DNA upon complex formation clearly complicates the docking exercise. These changes range from minor side-chain rearrangements to large protein backbone rearrangements and DNA helix deformations. A successful docking method should, therefore, be able to account for these.
2. How to locate the preferred interaction interface on the DNA and the protein? This searching/sampling of possible interaction interfaces is complicated by the homogeneous nature of the DNA helix and the conformational changes that are likely to occur upon complex formation. Aspects like a proper treatment of electrostatics and the plasticity in nucleotide/amino acid interactions deserve attention.
3. Depending on its efficiency to deal with the previous two aspects, a docking method it is likely to generate a large number of solutions with different

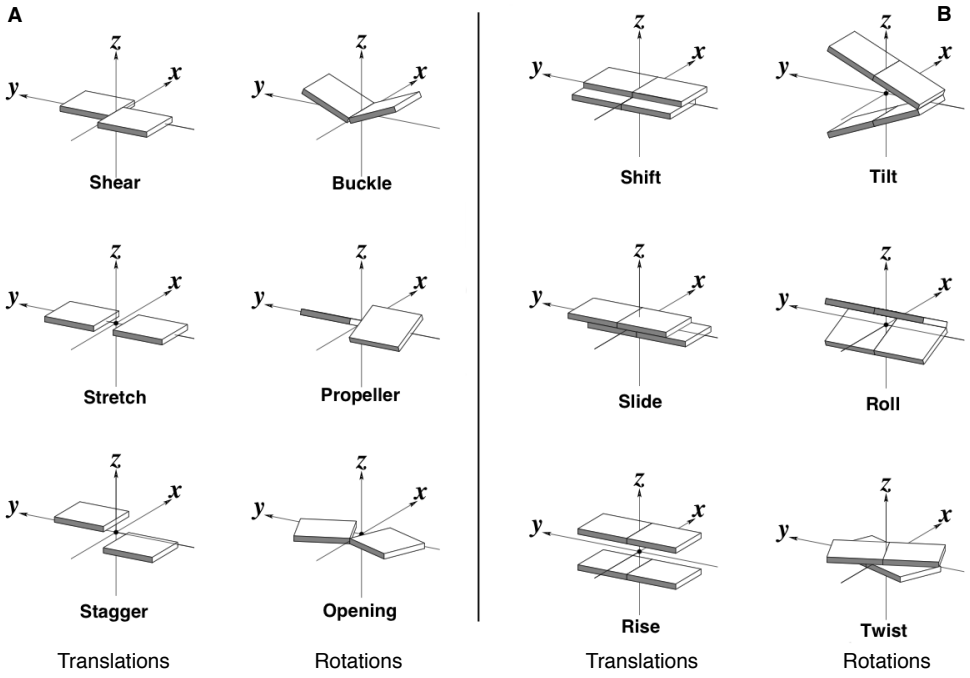



**Figure 2.3.** One block per base Calladin-Drew plot of the 6 base pair parameters (3 translational, 3 rotational, **A**) and six base pair step parameters (3 translational, 3 rotational, **B**) that together can unambiguously describe the conformation of the nucleic acid bases in a double stranded nucleic acid structure.

poses. An efficient scoring method should then be able to rank these such that the best scoring solutions reflect the correct protein-DNA interface(s). This exercise is complicated by pretty much the same reasons as the previous two challenges.

For any given docking method the above challenges affect several stages of the docking process: the selection of the starting structures to be used in the docking process, the rigid-body search and sampling through conformational space to identify the correct interaction interfaces(s), the treatment of flexibility and conformational changes and the final selection of most likely solutions (scoring and ranking). In the following sections we will discuss these stages in the context of the three main challenges formulated above.

## Starting structures

The preparation of every docking starts with the definition of the individual 3D structures of the biomolecules that need to be docked. They may exist in several conformations commonly categorized into three groups:

1. A "bound" conformation: Here, the biomolecules are (near) identical in sequence and conformation to that of their counterpart in the complex. These may reflect proteins that do not change their conformation upon binding to the DNA or a protein solved by NMR in the presence of DNA, effectively yielding the bound conformation. Furthermore, a bound-bound docking is often used to

validate the interface search/sampling stage of a docking method. This is essentially an exercise of separating the reference complex into its individual biomolecules and reconstructing it.

2. A "pseudo-bound" conformation: These represent molecules that are available in a bound conformation but with another molecule than the one in the complex to be modelled. This conformation may, however, reflect that of the biomolecule in the complex of interest more closely than an unbound conformation. Such structures may, for instance, be obtained by a homology modelling procedure.
3. An "unbound" conformation: These represent biomolecules in their free form, i.e. not bound to any other molecule, and thus possibly having a different conformation than that in the complex. The extent of this difference may vary widely as explained in the previous sections. An "unbound starting conformation" is the most common in practical docking settings.

It is common practise, nowadays, to solve the unbound structure of a protein using NMR or X-ray crystallography. When experimental structure are not available, homology modelling (113) can be used to generate 3D structural models. Furthermore, a number of experimental and computational methods are available that allow to probe the dynamics of structures, potentially generating models that more closely resemble the bound conformation (further discussed in "Dealing with flexibility").

In contrast to proteins, it is much less common to obtain the unbound DNA structure using experimental techniques. Due to the low proton density it is often difficult to obtain enough intra-molecular distances to unambiguously solve the structure using by NMR (reviewed in (123,137,346)). X-ray crystallography has less difficulty but, considering the likelihood of DNA conformational changes upon complex formation, it is not very relevant to solve first the unbound conformation. With the lack of an experimentally obtained structure one has to reside to computational modelling techniques to obtain a 3D starting structure. Here, the regular structure of the DNA double helix allows for easy computational modelling. In the following section, several DNA modelling approaches are discussed.

**DNA starting structures**

A proper choice of the DNA starting structure requires to answer two questions: 1) is the DNA sequence known and 2) what is the relevant DNA conformation to model? The DNA sequence involved in the interaction can easily be obtained using biochemical methods such as, for example, DNA footprinting (108). The situation of an unknown sequence is often related to the study of the effect of a variable base sequence on complex formation or the prediction of the interaction sequence on the DNA. The latter case imposes considerable difficulties as the protein-DNA interactions and the influence of the base composition on flexibility (protein-induced and intrinsic) are not known. In such cases all possible base pair combinations have to be evaluated, a computationally demanding exercise. Although this scenario is not the topic of this chapter, we will briefly discuss the two methods that try to deal with this problem.

Lafontaine *et al.* (170,171) created the program ADAPT to deal with this problem. The method is based on implementing a base sequence that can evolve during molecular mechanics simulations. A mean-field approach is used that is closely related to a number of so-called multi-copy algorithms (further discussed in "Implicit treatment of flexibility"). This method was shown to converge to the near native DNA sequence in most cases. When a slightly curved starting conformation is used, bases that are known

to facilitate the curvature are favoured ($A_nT_n$ tracts separated by GC pairs). Another method was developed by Cambell *et al.* (47) who docked the bound protein of five different transcription factors to model-built DNA. The method uses an optimized pattern recognition approach based on geometric hashing, in combination with Monte Carlo (MC) minimization to test all possible DNA sequences to which the proteins could bind. The selection is based on an energy function using a hydrogen-bonding interaction term. This method was able to identify the proper sequence in most cases. Its application is, however, limited to the bound conformation of the protein and the resulting DNA conformations are often distorted.

For the most frequent situation in which the DNA sequence is known the influence of the base composition on the conformation is no longer a variable. However, predicting the unbound conformation of the DNA and/or the change in conformation upon complex formation is not trivial (further discussed in "Dealing with flexibility"). With this in mind most docking methods start from an idealised canonical DNA conformation. Such 3D structural models can be easily generated by a number of software suites (Table 2.1) using regular nucleotide building blocks with parameters often derived from fiber diffraction studies (50). As most of the biologically relevant DNA molecules are found in the B-type conformation, the choice for this conformation as starting point for the modelling procedure seems justified, even if the final conformation in the complex might adopts a state in between A- and B-DNA (45,230,236,289,314).

**Table 2.1.** Overview of various DNA modelling software.

| Method | Description |
|---|---|
| NAMOT (316) | Builds regular A- and B-form DNA-models. Provides control over helix bending and phasing as well as base pair (step) parameters. Stand-alone software freely available at: http://namot.lanl.gov |
| NAB (201) | Stand-alone software implemented as scripting language. Builds regular helices and complex structures such as triplex- and tetraplex DNA. Implements an AMBER implementation for force-field calculations. Freely available at: http://casegroup.rutgers.edu<br>Reduced functionality available as web server: http://structure.usc.edu/make-na |
| 3DNA (194,195,356) | Builds regular A-, B- or C-form DNA-models. Allows rebuilding from a table describing the conformation of the helix as a collection of base pair (step) parameters. Stand-alone software freely available at: http://rutchem.rutgers.edu/~xiangjun/3DNA<br>Reduced functionality available as web server: http://w3dna.rutgers.edu |
| MDDNA (85) | Builds structural models from a database of tetranucleotide steps derived from MD simulation. Uses 3DNA as structure building engine. Accessible as a web server: http://humphry.chem.wesleyan.edu/MDDNA |
| DNAtools (225,332) | Builds regular A- or B-form DNA-models or bent conformations by a set of structural parameters. Allows for energy minimization of the generated models. Accessible as web server: http://hydra.icgeb.trieste.it/~kristian/dna |
| CURVES/ JUMNA (178,179) | The DNA analysis program CURVES can be used in combination with JUMNA a program for the energy optimization of nucleic acid structures to build a variety of complex DNA structural models. CURVES can be freely obtained at: http://gbio-pbil.ibcp.fr/Curves_plus |

As mentioned above, the conformation of unbound DNA has been proposed to play a role in recognition and complex formation (indirect readout mechanism). Consequently, starting from a 3D structural model that more closely resembles the true conformation of unbound DNA might be beneficial with respect to locating the correct interaction interface(s). A number of methods have been proposed to predict the conformation of the DNA for a given sequence. The MDDNA (85) web server generates 3D models by using a database of MD simulation results on all 136 unique permutations that can be made with tetranucleotide steps (33,84). Rohs *et al.* used a Monte Carlo (MC) simulation of the Papillomavirus E2-DNA binding sites to predict the intrinsic DNA bending (263). Farwer *et al.* (100) took a different approach: they used an energy function to predict the conformation of the bases and of the backbone and build an all atom 3D structural model. An earlier attempt was made by Breslauer *et al.* (40) who generated a library for all 10 Watson-Crick nearest neighbour interactions based on thermodynamic data obtained from calorimetric experiments: they proposed that their tool should be useful in the generation of all-atom DNA models reflecting sequence-dependent stacking effects. Apart from these prediction methods, molecular dynamics simulations starting from ideal canonical B-DNA can also be used as a tool to generate an ensemble of conformations that reflect more closely that of the free DNA structure in solution (32,55,233).

Apart from Aloy *et al.* (8) who used an energy optimized B-DNA structure, all current docking methods often use idealised canonical DNA conformations as starting structure and might implement some treatment of flexibility during search/sampling to account for conformational changes taking place in the DNA.

## Searching conformational space, locating the protein-DNA interface

Once the starting structures have been obtained the generation of plausible solutions by docking is an exercise of sampling conformational space. The goal of sampling is the systematic matching of the surfaces of the biomolecules in question in search of the proper interaction interface(s) using a number of criteria. The amount of space that needs to be sampled depends on the size and number of individual biomolecules, the procedure of selecting favourable binding interfaces, the expected conformation of the biomolecules in the complex and the treatment of flexibility.

In our general introduction about protein-DNA complexes we emphasized the complexity of these systems with respect to their mode of interaction and the possible conformational changes occurring both in the protein and DNA during complex formation. A docking method that is able to deal with all these challenges has not yet been developed, and, as a consequence, many methods address either a part of the docking problem or focus on certain classes of protein-DNA complexes. Simplifications often involve the reconstruction of the complex from the individual biomolecules in their bound state, focussing on the art of finding the correct interaction interface(s), or the choice of a protein-DNA system for which conformational changes and size are limited. There are a number of transcription factors that are good candidates to match these criteria. Examples are the bacteriophage 434 Cro protein (222), *E. coli Lac* repressor (264) and bacteriophage P22 Arc repressor (257). They use well-known single DNA binding motifs such as Helix-Turn-Helix or β-sheet to interact with the DNA major groove. For these, typically, quite a number of high-resolution crystal structures are available and the subtle variations often found in the operator sequences provide a challenge for methods

aiming at locating the interaction interface. Many of these transcription factors, however, interact with the DNA as dimers. Multiple interfaces and conformational changes make these three-component systems considerably more difficult to dock than their monomeric counterparts. As such, many docking methods were developed using only a monomer of the protein and one operator half-site, or docking the dimer as one structure. Because of the limited conformational changes in these systems, they were often treated as rigid bodies.

Although many of the current protein-DNA docking methods only make use of rigid-body docking techniques, even those that include flexibility almost always start with a rigid body docking stage to generate the protein-DNA encounter complex. Therefore, we start with an overview of the various rigid-body docking techniques and then discuss the treatment of flexibility.

**Rigid-body docking**

In rigid-body docking, the search through conformational space is reduced to six degrees of freedom: three translations en three rotations. Several methods have been developed that sample conformational space in a rigid body manner (Table 2.2). Sampling using the Fast Fourier Transform (FFT) algorithm (152), as implemented for example in ZDOCK, FTDOCK and DOT have been used in a number of cases (1,8,98,99,262). FFT is a fast search algorithm allowing for an exhaustive sampling of conformational space. Protein and DNA are represented as one or more values at every point on a discrete grid. Usually the DNA is kept fixed while the protein is allowed to rotate and translate. Docking is then performed by overlaying the grids (effectively done by convolution in Fourier space) where an overlap of the surface points is favoured while an overlap of the molecules interiors is penalized. Using FFT, all possible translations can be calculated at once for a given rotation. The procedure is repeated for all possible (or a subset of) rotation angles. The docking program ESCHER (15,73) compares the solvent accessible surface of the target and probe in all possible directions by preparing parallel slices of the surfaces and representing them as polygons; the 'goodness of fit' between polygons is subsequently evaluated. Liu *et al.* (191), Bastard *et al.* (20) and Knegtel *et al.* (155,156) used rigid-body Monte Carlo minimization. Finally, Poulain *et al.* (252) used a course-grained approach implemented in ATTRACT (355) for the modelling of protein-DNA systems. Coarse grained approaches aim at reducing the number of particles in a simulation thereby reducing the computational time or increasing the number of sampling steps per unit of time. Coarse graining should, however, be done without loosing too much of the characteristics of the molecules such as shape, steric and electrostatic properties.

Some of the rigid body docking methods were tested for their ability to predict loss of binding by mutations of protein interface residues (99) or to retrieve the correct DNA sequence from a larger pool of mutated sequences (252). These attempts worked rather well if the bound form of the protein was used since the residues at the interface have their side chains in a conformation that favours interaction. The reduced number of intermolecular contacts upon mutation of these residues was brought forwards as an indicator of loss of binding affinity. Similarly, monitoring the loss of intermolecular contacts upon mutation of the base sequence allowed identifying the near native DNA sequences from the total pool of mutated sequences. Both of these methods, however, performed less well when unbound proteins were used stressing the importance of side-chain flexibility in docking.

For many of the above-mentioned methods it is difficult to impossible to dock more than two molecules simultaneously. This is one of

the reasons why many only use the protein monomer and one operator half-site as test systems or consider a dimer as a single unit. This, however, requires knowledge of the conformation of the protein dimer and might limit an effective treatment of flexibility to account for independent motions of the protein monomers.

The discussed rigid-body docking methods will not result in a single solution but rather in a collection of typically hundreds to several thousands of solutions. The choice for the correct model or ensemble of models is not trivial. Here, it is the task of the scoring function to rank all solutions by combining different scoring parameters and filter steps. These will be discussed in the following section.

**Scoring functions**

The various docking methods use different scoring algorithms to rank the generated solutions. The choice of a scoring method is very much system-dependent. The use of a desolvation term for instance, has no meaning in a membrane system and a nucleotide/amino acid pairing potential is of no use in a protein-protein docking setting. Therefore, a universal and fast scoring function that is able to accurately distinguish between native and non-native docked conformations, has not yet been devised. The development of an efficient scoring function is a science by itself. An in depth

**Table 2.2.** Overview of current protein-DNA docking methods with information on the starting structures used, the algorithm for rigid-body optimization and the treatment of flexibility.

| Method | Starting structures | | Rigid-body | Flexible docking |
|---|---|---|---|---|
| | protein | DNA | Optimization | |
| FTDOCK (8) | unbound | B-DNA, minimized in JUMNA | FFT | soft-body |
| ATTRACT (252) | bound | B-DNA coarse grained representation | quasi-Newton minimizer | - |
| DOT (1,98,262) | bound | B-DNA (201) | FFT | soft-body |
| ESCHER (15,73) | unbound | B-DNA | Polygon matching | - |
| Liu *et al.* (191) | pseudo-bound | homology modelled | MC | - |
| MONTY (155,156) | unbound | B-DNA | MC | Side-chain MC |
| ZDOCK (99) | bound, bound mutated | bound | FFT | - |
| MC2 (20) | bound with loop ensemble | B-DNA | MC | Flexible loops by mean-field multi-copy |
| Sandmann *et al.* (270) | bound | Bend B-DNA models | Molecular Dynamics: DNA models were shifted towards the protein in steps. Every step was finished with an energy minimization for protein side-chains. Final minimization on DNA and protein main-chain/side-chain in steps for best models | |

Use abbreviations: MC; Monte Carlo minimization, FFT; Fast Fourier Transform

discussion of scoring functions is described in the review of Halperin *et al.* (120). Here, we focus on some elements specific to protein-DNA complexes and describe their use in the current protein-DNA docking methods (Table 3.3).

All docking methods use a set of physicochemical energy terms as part of their scoring function. These include non-bonded terms such as electrostatics and Van der Waals energies and often a solvation screen (99,331). They are often defined as 'binding energy' calculated as the total energy of the system minus the energy of the individual free components or 'interaction energy' considering only intermolecular energy terms. Electrostatic complementarity is used by all methods to favour DNA interaction motifs on the protein over other patches, due to their above average polar and positively charged amino acid composition.

The rigid-body search algorithms rely heavily on geometric criteria to favour a large buried surface area. This is a powerful constraint for many transcription factor - DNA complexes. Most of them interact with the major groove and do so with well-defined structural motifs. This type of interaction generates a 'snug fit' with a significantly higher buried surface area than any other conformation.

Many methods implement a hydrogen bonding term due to their frequent occurrence in protein-DNA complexes. Such a term can favour solutions based on the number of formed hydrogen and is sometimes biased to favour amino acid – base contacts.

Structure quality assessment is also an important aspect of the final selection process. In contrast to protein structure validation software such as WATCHECK (131), PROCHECK (176), MOLPROBITY (70), PSVS (34) and CING (http://nmr.cmbi.ru.nl/cing), which are now widely used, "true" nucleic acid structure validation software is currently virtually non-existent. DNA structural analysis is limited to programs that determine a set of structural descriptors (179,194,195). These are subsequently used to identify abnormalities but this is not a true validation.

The *ab initio* methods that use scoring functions based purely on physicochemical

**Table 2.3.** Scoring schemes used by the various protein-DNA docking methods

| Method | Scoring scheme used |
|---|---|
| FTDOCK (8) | Solutions are generated with consideration of shape- and electrostatic complementarity. Subsequent geometric filter to remove docking artefacts, a knowledge based filter for specific base recognition and a final nucleotide/amino acid pairing potential |
| ATTRACT (252) | Shape complementarity and minimization of the interaction energy |
| DOT (1,98,262) | Shape- and electrostatic complementarity. Minimization of the interaction energy. |
| ESCHER (15,73) | Shape- and electrostatic complementarity. Hydrogen bond term favouring nucleotide base/amino acid contacts. |
| Liu *et al.* (191) | Energy function composed of a nucleotide/amino acid pairing function, a Van der Waals packing term and a knowledge based positional restraint energy. |
| MONTY (155,156) | Energy function composed of a hydrogen bond term favouring nucleotide base/amino acid contacts, a Van der Waals term and a knowledge based energy term. |
| ZDOCK (99) | Shape- and electrostatic complementarity, desolvation |
| Sandmann *et al.* (270) | Electrostatic complementarity and minimization of the interaction energy. |
| MC2 (20) | Minimization of the interaction energy. |

parameters tend to be insensitive to several system specific features like the conformation of the molecules and intermolecular contacts that are generally made. As a consequence, these methods perform best for small systems with well-defined interaction interfaces and little conformational change upon binding (such as the transcription factor – DNA complexes described above). However, when the size of the interface starts to increase or the protein interacts with multiple or different DNA binding motifs than the ones commonly used in the transcription factor test systems, the docking performance decreases. Such cases often involve proteins that interact with the minor or major groove, or both, and possibly at multiple sites. The physicochemical properties along the DNA also pose a challenge for scoring: they are often too uniform to unambiguously predict the correct complex, which often leads to false positive solutions. These can be rotational false positives in which the protein is rotated by 180° in the DNA groove, frame-shifted conformations in which the protein is shifted by one or more base-pairs up- or downstream from the correct operator sequence or cases in which the key protein residues do interact with the bases but the overall orientation on the DNA is different (8,155).

To deal with these problems many docking methods nowadays use a wide range of additional information about the complex under study to eliminate a large number of possible outcomes. Experimental data, for instance, have been used to favour solutions that satisfy experimentally observed contacts (8,155,156,191). The benefits of using additional information have been shown numerous times in protein-protein docking (reviewed in (323)). The same principles can be used in protein-DNA docking. In the following section, the available sources of experimental data that can be used for protein-DNA docking are discussed, followed by a discussion of other information sources useful in case experimental data are not available in sufficient quality and/or quantity.

**Experimental information**

Biochemical and biophysical experiments are widely used to gain insight into biomolecular interactions. They can provide information about the recognized DNA sequence and specific nucleotides or amino acids involved in the interaction. The large number of useful biochemical and biophysical experimental methods has been reviewed in the past (216,323). Here, we focus on some aspects specific for protein-DNA docking.

Mutation data have been one of the most used source of information. The general idea is that the chemical modification of a critical interface residue weakens or completely abolishes the interaction with the partner molecule while modifications of non-interface residues have (should have) no effect. For mapping the interface, the mutations should only affect the solvent accessible residues. Mutation of a buried residue can still inactivate the molecule due to a loss in structural stability. Furthermore, the mutation itself should not enhance the interaction with the target molecule. In-depth and systematic mutagenesis studies can be performed using the alanine-scanning approach (75,223,224) or by specifically targeting residues based, for instance, on their conservation. Results of various alanine-scanning experiments have been conveniently gathered in a web-database (http://www.asedb.org, (310)). Although this database contains a limited set of protein-DNA systems it is bound to increase over time.

Mapping the interaction interface on the DNA can only in part be investigated by base mutation studies. Furthermore, interpretation of DNA mutation data requires careful considerations. Protein-DNA contacts are not necessarily base-

**Table 2.4.** Examples of docking methods that have used additional information in the docking procedure.

| Complex | Information used | Reference |
|---|---|---|
| *FTDOCK* | *Used as filter* | |
| 8 monomeric repressors | FP, EP | (8) |
| *Molecular Dynamics* | *Distance Restraint* | |
| Fis | NC, EthI, MthI, EC | (317) |
| *Molecular Mechanics* | *Used as filter* | |
| Pf3-ssDNa | CSP | (103) |
| *MONTY* | *Part of the scoring function* | |
| CylR2-DNA | CSP | (266) |
| 434-Cro and Lac | EthI | (156) |
| LexA | EthI | (157) |
| Fur | EP | (191) |

Used abbreviations: CSP; Chemical Shift Perturbation data, EthI; Ethylation Interference, MthI; Methylation Interference, FP; DNA footprinting data, EP; Emperical protein-nucleic acid pair potential, NC; Chemical modification turns residue into site-specifc nuclease cleavage, EC; Evolutionary conservation

specific but can be targeted to either pyrimidine (T-C) or purine (A-G) base classes. Furthermore, some base pairs may facilitate a conformational change in the DNA that is important for the interaction while not making any contact with the protein (indirect readout); mutating such bases can result in a reduction of the binding affinity even if they are not directly involved in intermolecular contacts. Similar problems also affect the reliability of double mutant cycles (48) in which a set of mutants is created for both molecules of a complex.

DNA backbone binding can be studied by ethylation interference. In this technique the phosphodiester groups on the backbone are at random chemically labelled using ethylnitrosourea (EtNU) yielding phosphotriester groups. Distinguishing between DNA molecules that are able to bind a protein with high affinity and those for which the ethylation has lowered the affinity allows for the mapping of the phosphate groups contacted by the protein. Labelling using dimethylsulfate is also commonly used to probe contact with purines, in particular G.

Some DNA-binding proteins can be transformed into site-specific nucleases when specific amino acids residues are mutated into cysteines and conjugated with synthetic nucleolytic agents (21,247). The modified protein is than able to cleave DNA through an oxidative attack at the C-1'H of particular nucleotides. The DNA scission reaction thus provides information about the proximity between the mutated protein residue and the nicked DNA site.

The minimum required DNA sequence length for interaction can be rapidly identified using DNA footprinting methods (108). DNase-I is often used to cut the DNA that is not protected by protein binding. Isolation and sequencing of the bound DNA afterwards provides the interaction sequence, which provides valuable information for the generation of DNA structural models for docking. This does not provide, however, any information on base-specific recognition. Furthermore, since DNase-I is a fairly large protein that needs to access the phosphodiester bonds the footprinted DNA is therefore bound to be larger than the minimum required interaction sequence.

Apart from biochemical methods several biophysical methods also provide useful information. Hydrogen/Deuterium exchange detected by mass spectrometry (121) (or

NMR) provides residue accessibility data which can give indicate the location of a residue at an interface (174). NMR can also readily be used to map interactions by measuring Chemical Shift Perturbations (CSP) in titration like experiments (196). However, considering the difficulty of assigning DNA and the large-scale conformational changes often observed, CSP data for DNA are difficult to obtain and interpret. Reliable CSP data can, however, easily be obtained for the protein side. H/D exchange, cross-saturation and saturation transfer experiments can provide similar information. Cross-saturation data are believed to be more reliable than CSP as the latter can suffer from false positives due to (remote) conformational changes. Next to these, Redisual Dipolar Couplings (RDC) can be used to provide information on the relative orientation of the molecules (reviewed in (123,137,346)). An overview of the information sources that have been used in various docking methods is given in table 2.4.

Many docking methods include information in a filtering step to select the final solutions: solutions are selected if their interfaces are aligned, if specific residues are at the interface or if certain interface residue pairs are found in close vicinity. Other methods include external information as an additional energy term in the scoring function. MONTY (156) for instance, rewards as an energy bonus during the docking, the contacts that are in agreement with the experimental data. Tzou *et al.* (317) used various sources of experimental data to map the DNA interface and included this information as a set of distance restraints to steer their molecular dynamics docking approach.

**Database-derived information**

The number of protein-DNA complexes deposited in the RCSB protein databank is increasing steadily. With currently over 1400 complexes the database has been surveyed many times using statistical methods to study various aspects of protein-DNA recognition. The results are often made available via user-friendly web interfaces such as AANT (129) and ProNuc (10). These give researchers easy access to a wealth of information. A number of these studies focused on elucidating the much pursued protein-DNA interaction code (25,59). Although, as discussed previously, this code appears to be degenerate in both directions the studies did reveal some common trends. From these, amino acid to base interaction potentials could be derived that have been used as energy term during docking or as filter after docking (8,190,320,351).

One of the first empirical nucleotide/amino acid pairing potentials (199) was developed based on a number of protein-DNA complexes containing zinc finger domains. This potential described the amino acid to base interactions in these types of complexes but was not able to do so for many transcription factor-DNA complexes. To deal with these shortcomings, Aloy *et al.* (8) developed an empirical score for nucleotide/amino acid pairing based on transcription factor DNA complexes and used it as a filter to select the final models. Liu *et al.* (191) used a knowledge-based potential (190) as energy function in their docking method. This energy function differs from others in that it uses nucleotide triplets as interaction unit instead of only one nucleotide. As such it also takes into account bi-dented amino acid to base interactions as well as DNA structural deformation and the contribution of local sequence/structure interactions (indirect readout).

The structural and physical properties of DNA provide important constraints on the binding sites on the surface of DNA-binding proteins. Their characteristics may be used for predicting DNA-binding sites on unbound proteins, which, again,

can be useful to either drive the docking or filter the solutions. The currently available software programs that aim at predicting DNA binding sites on proteins come in two flavours: those software that require only the amino acid sequence for the prediction (3,5,136,166,168,234,339) and those that also require the 3D-structure of the protein (143,311). Many of these are available as web servers (3,5,136,143,166,234,311,339). The first approach has as benefit that the structure of the protein in question is not needed to make a prediction as they are mainly driven by evolutionary conservation of residues or residue motifs and the propensity for a given residue to be located at the interface. Other studies, however, have shown that the inclusion of structure related data such as solvent accessibility, charge, dipole/quadrupole moments (4), average electrostatic potential, secondary structure, crystallographic B-factors, protein surface and neighbouring residues (3) and location in a cationic patch (130,298) improve the prediction results significantly. This, however, requires the 3D-structure to be known, but, in a docking setting, this is often the case. Some prediction methods implement a probalistic recognition code between amino acids and bases (205). This approach does, however, have several drawbacks (219,244): the developed models tends to be most successful for a given family of DNA binding proteins and they often assume a single binding mode (24). In addition, the stabilizing influence of long-range interactions is often difficult to include in such methods.

When a prediction yields a possible DNA binding interface, no assumption is however made about the function of the protein patch. Assigning a function is one step further. Methods have been developed that aim at exploiting the knowledge of the three-dimensional structure of proteins to establish whether functional inheritance within a given protein super family is likely to be valid and can be used for a prediction. Often information about the functional sites of related proteins is obtained from databases such as PFAM (101), SMART (185), SCOP (226) or PROSITE (135). The relationship between sequence similarity and functional similarity is however weak (78,127,267,347); the same catalytically active residue might have different functions in different catalytic sites. One such method for the prediction of protein functional sites was used in a protein-DNA docking study using FTDOCK (9). Here the prediction data were used as a filter in the selection of docking results. The method competed favourably with the manual introduction of biochemical information as filter. Next to the advances have that been made in the development of prediction methods, Gao *et al.* (112) showed that protein-DNA docking can be a valuable tool for the identification of the DNA-interacting interface on a protein. By docking both the bound and free form of 44 different DNA binding proteins to a stretch of non-specific canonical B-DNA they found that all 44 proteins bind to the non-specific DNA using the same interaction sites as they would use to bind to specific DNA sites. These sites were favoured over non-DNA-interacting surface patches on the proteins.

Algorithms aimed at predicting binding DNA sequences in genomes have seen a more rapid development in the last years. Various tools/methods have been proposed including the use of consensus sequence (71), weight matrices, information content and protein-DNA recognition patterns (158,273,300). The problem of predicting a target sequence in a large pool of possibilities depends on the type of DNA-interacting protein. In case of enzymes the sequence is rather well defined: A restriction enzyme often interacts (cleaves) the sequence it is designed to cut with only little consensus requirements for the stretch where the specific sequence is embedded in. A transcription factor on the other hand might interact with

various, slightly different sequences. From a biological point of view this is desirable as it provides a way to control transcription. From a prediction perspective it is, however, a difficult case. Docking can be used as a tool to select the most favourable interaction sequence from the pool of possible sequences a given transcription factor can interact with (252).

## Dealing with flexibility

In the previous section we introduced the search through conformational space for the correct binding interface by considering the biomolecules as rigid entities and pointed to the use of additional information to limit this tedious search. Considering the system as rigid, however, is a gross simplification. As discussed before, proteins and DNA often change their conformation upon complex formation, sometimes quite drastically. Ignoring flexibility is therefore likely to generate a biologically irrelevant model. Including flexibility in the docking process, however, is far from easy and constitutes one of the major obstacles in the development of effective docking methods to date.

In rigid-body docking, the search through conformational space is restricted to six degrees of freedom, three translations and three rotations. When flexibility is considered the number of degrees of freedom increases dramatically. Allowing more flexibility is not only computationally demanding it also requires a close guarding of the quality of the generated results, especially for DNA.

In this section we will discuss how current docking methods deal with flexibility. Roughly, one can separate them in methods that include flexibility explicitly and those that do so implicitly. In implicit methods the reorientation of residues is not an active process during the docking. These include ensemble docking and soft body docking. The explicit methods allow for conformational changes directly in the docking procedure by letting side-chain and/or main-chain atoms of residues move in a series of molecular dynamics/mechanics or Monte Carlo simulations.

### Implicit treatment of flexibility

With implicit flexibility an approximation of flexibility is made either by the use of a soft-body approach or by means of an ensemble of starting structures representing the different conformational states of the components that need to be docked.

In a soft-body approach, the residues at the interface are allowed to interpenetrate each other, approximating the side-chain rearrangement often found upon complex formation. Because interpenetration is likely to generate severe steric clashes the final solutions need refinement to remove them. This method is only able to deal with side-chain rearrangements although final refinement can result in small backbone rearrangements. Soft-body docking is often performed by reducing the electrostatic repulsion and allowing some overlap of the atoms (8). Such approaches are also used to promote interaction of the protein with the DNA bases by reducing electrostatic repulsions: this is often done by scaling down the electrostatic energy term or by reducing the partial charge of the phosphate groups and increasing that of the chemical groups in helix grooves (8).

In ensemble- or multiconformer docking, the docking is performed with all models in an ensemble either one-by-one (cross-docking), a process that considerably increases computational time, or all at once by means of a mean-field approach. Various sources can be used for the ensemble. The most obvious one is an NMR ensemble, but also snapshots from a molecular dynamics/mechanics run, models from homology modelling or models obtained from a normal mode analysis can be used. They have been applied to proteins and DNA in different ways.

*Protein ensembles*

A protein ensemble can be constructed to represent flexibility on several levels ranging from side-chain rearrangements to rigid domain movements. Ensembles can be used to represent different loop conformations of a protein that are often interacting with the DNA. Such an approach was used by Bastard *et al.* (20) in their MC2 docking program to dock the crystal structure of the bound form of the Drosophila prd paired domain to its DNA target both in the bound and canonical B-DNA conformations. In their mean-field method, multiple copies of the DNA interacting loop were used with the same initial weight. During iterations in the docking run, the weights of the conformations change according to a Boltzmann criterion: the docking partner and the remaining protein interact with the average field created by the ensemble of loop copies. Each copy interacted with the rest of the protein and the partners, but does not see the other copies. Using iterative Monte Carlo cycles the method usually converges to a single conformation, the one with the highest weight. This approach was successful in recovering the proper loop conformation from the many possible ones. Its application is, however, limited to cases where the DNA binding interface of the protein is at least partially determined from experimental information or by prior rigid-body docking.

Essential dynamics can also be used to generate a set of realistic starting structures representing possible conformations of the protein (72,322). Here, the main flexible degrees of freedom in a set of protein conformations is captured by a number of vectors determined in a principal component analysis (PCA). The deviations between the atom coordinates in the different conformations, usually obtained by MD, are used to create a square covariance matrix. Eigenvectors obtained by diagonalization of this matrix represent the main fluctuations in the protein and the eigenvalues the amplitudes of the fluctuations.

Normal Mode Analysis (NMA) is a statistical technique that describes the motion in a protein as a set of basic vectors (normal modes). Each vector describes a certain movement and any conformational change can be expressed as a linear combination of these vectors. The coefficient of a normal mode represents its amplitude. The normal modes describe a continuous motion around a single equilibrium conformation. Theoretically, this model does not apply to systems that exist in several conformational states. However, in practice, normal modes seem to predict rather well conformational changes observed between bound and unbound protein structures and have been used as such extensively (18,58,76,119,214,302,309).

Another advantage of the normal modes analysis is that it can discriminate between low and high frequency modes. The low frequency modes usually describe the large-scale motions of the protein. It has been shown that the first few normal modes, with the lowest frequencies, can already describe most of the conformational changes. This allows reducing the degrees of freedom considerably while preserving the information about the main characteristics of the motion. Therefore, many studies use a subset of the lowest frequency modes for analyzing the flexibility of proteins. The normal modes can further be used for predicting hinge-bending movements, for generating an ensemble of discrete conformations and for estimating the protein's deformation energy resulting from a conformational change. The higher frequency modes on the other hand are suitable for detecting conformational changes in loops. One drawback, however, is that no information is provided about neither the direction in which the conformational change is likely to take place nor the amplitude of the change (reviewed in (18,265)).

Rigid domains linked to each other by flexible hinges are often seen in DNA-interacting proteins. Upon complex formation these rigid domains might rearrange using their flexible hinge to form the final complex. This process can be modelled by: 1) identifying the hinge points, 2) cutting the molecules between the domains and 3) docking them as separate (rigid) bodies. FlexDock (290,291) in combination with the HingeProt algorithm (96) has been successfully used in this way (in a combination of two-body docking runs), to model the large conformational changes occurring in Replication Protein A upon interaction with its single strand DNA target.

*DNA ensembles*

An ensemble representing unbound DNA (discussed in "DNA starting structures") might be useful to mimic the indirect readout mechanism of encounter complex formation but will likely be of little use in predicting the DNA conformational change brought about in transition to the final complex. Instead, an ensemble of pre-bend and twisted DNA conformations, more closely representing the bound conformation, can be used. The most straightforward implementation of this approach is the generation of evenly bent and twisted DNA 3D structural models. These can be made by stand-alone software programs such as NAMOT (316) and NAB (201). Information about the bend angle to be modelled can, for instance, be obtained from biochemical methods such as a gel-shift assay (200) or from the topology of the DNA binding interface on the protein as used by Sandmann *et al* (270).

More accurate DNA models were obtained by Liu *et al*. (191) who used structural similarities among related transcription factor DNA complexes in a homology modelling approach. Here, all the DNA conformations in the 141 different transcription factor DNA complexes in the RCSB databank were superimposed using their backbone heavy atoms followed by clustering. The structures in the clusters were then used to build 3D models for a given DNA sequence. Molecular dynamics simulations could be used to generate more accurate bent DNA structures, provided some information about the orientation of the bend is available (65,66).

**Explicit treatment of flexibility**

Explicit methods do allow side-chain and or backbone rearrangements during the docking by including molecular dynamic/ mechanics stages or after docking as a final refinement step using molecular dynamics. Side-chain rearrangements are almost always observed upon complex formation but only a small number of methods account for them by including a refinement stage in which side-chain orientations are optimised (20,155,156,270). To reduce the search through conformational space many methods use rotamer libraries. These libraries are derived from a statistical analysis of side-chain conformations in high-resolution crystal structures. They are often backbone-dependent meaning that the side-chain conformations depend on the backbone conformation (89).

The MONTY approach developed by Knegtel *et al.* (156), although initially a rigid-body method (155), was refined to be the first method to incorporate implicit side-chain flexibility in a Monte-Carlo minimization procedure. It was successfully applied to model the half site of the 434 Cro, Lac and Gal complexes. Introducing DNA flexibility and experimental restraints proved to overcome some of the difficulties associated with docking unbound components. The method, however, did require the components to be close together in almost the true conformation otherwise the DNA would tend to curl around the protein.

Sandmann *et al.* (270) used pre-bent DNA models. These were selected by preliminary electrostatic calculations to assess

whether the DNA could be pre-oriented electrostatically in the potential of the protein and if there exists an electrostatic recognition for different DNA curvatures. The resulting rigid-body docked solutions were further refined using molecular dynamics.

Tzou *et al.* (317) docked three helix-turn-helix proteins (CAP, Rep, Fis) to their respective DNA targets using knowledge-based distance restraints in a series of molecular mechanics and dynamics simulations. They used unbent canonical DNA as starting model and docked it to the protein using experimentally derived distance restraints. These included knowledge-based restraints derived from ethylating interference, DNA-interacting amino acids derived from structure analysis and mutation data. The DNA was treated as flexible but interproton distance restraints were defined for all distances under 5 Å in B-DNA. This method proved to be efficient in allowing the DNA to adapt its conformation to that of the protein without gross loss of B-DNA conformation. This was the second method after that of Knegtel *et al.* (156) that effectively incorporated DNA flexibility during docking.

Altogether, there have not been many methods that accounted for flexibility explicitly. Often, the docking methods have been developed with the focus on reconstructing the correct encounter complex in a rigid-body manner. Methods that tried to account for all-atom flexibility were often faced with DNA deformations. These could potentially be prevented by "freezing" a number of degrees of freedom during the simulation (169) such as planarity of the base ring and the sugar pucker or by treating the nucleotides as individual units that have little influence on each other. This, however, prevents the accurate simulation of base pair stacking.

In the last couple of years, the trend shifted to a "multi-stage" approach in which the strong aspects of several methods are combined to obtain the desired results. For example, a rigid-body approach can be used to predict the initial encounter complex while a molecular dynamics simulation is used afterwards to account for flexibility and refine the complex. Molecular dynamics simulations themselves are not the best choice for performing docking due to their computational requirements allowing to probe flexibility only on the nanosecond time scale; micro- to millisecond timescale motions, which often occur in protein-DNA systems, remain out of reach. Modern molecular dynamics methods and force fields are however able to accurately simulate base pair stacking accounting for the shortcomings of the docking methods. This "multi-stage" approach will, however, only works if the docking solutions are close enough to the "true" complex.

Lindahl and Delarue (187) have used normal mode analysis to account for the problem described above. In this approach the initial docking results were refined by minimizing the interaction energy in a complex along 5–10 of the lowest frequency normal modes directions, resulting in an improvement of the "steric fit" between proteins and DNA: the degrees of freedom in the search space are the amplitudes of the normal modes. This method is able to sample more conformational space than regular molecular dynamics simulations, requiring a less accurate solution from the docking simulation. This approach is available through a web portal (http://lorentz.immstr.pasteur.fr).

Also Zacharias and Sklenar have used harmonic modes to describe protein flexibility allowing for relaxation of these modes during docking to improve the steric fit between ligands and the minor groove of the DNA (354).

Roberts *et al.* (262) suggested to start the docking using a canonical B-DNA model, implement flexibility in the docking and

analyze the generated models in search of trends in conformational changes taking place in the DNA. This information can subsequently be used to generate an ensemble of pre-bend DNA models for a subsequent docking round.

## Extending HADDOCK to protein-DNA systems

The challenges of assembling the correct interaction interface and dealing with conformational changes are not exclusive to protein-DNA docking but are applicable, in a variable extend, to the docking field in general. Dealing with these challenges is the focus of many methods including the data-driven docking method HADDOCK (High Ambiguity Driven DOCKing, (74,86)) developed in our group. With respect to other methods, HADDOCK has several unique features that makes it especially suitable to deal with these challenges. Two of these deserve special attention:

1. HADDOCK can make use of a wide range of biochemical and biophysical data sources that can provide information about the interfaces of biomolecular complexes. While most methods are merely using this information as a filter to select the best solutions, HADDOCK was designed to actively use the information to drive the docking process, thereby limiting the conformational search.
2. HADDOCK allows for various degrees of flexibility that enable the biomolecules to adapt their conformation during complex formation. These include implicit flexibility by means of structural ensembles and explicit side-chain and backbone flexibility by means of simulated annealing and molecular dynamics stages.

HADDOCK uses CNS (41) as structure calculation engine. This is the same software that is commonly used to calculate structures from data derived from NMR spectroscopy and X-ray crystallography experiments. CNS provides native support for nucleic acids and can easily be extended to cover other types of (bio)molecules by providing the proper topology and parameter files. HADDOCK's unique features together with its native support for nucleic acids should make it a promising method to deal with the challenges of protein-DNA docking as described above. The next sections discuss this in more detail.

### The use of experimental data

HADDOCK encodes the experimental (or predicted) information in Ambiguous Interaction Restraints (AIRs) to drive the docking. These are similar to the ambiguous distance restraints commonly used in NMR structure calculations (232). An AIR defines that a residue on the surface of a biomolecule should be in close vicinity to another residue or group of residues on the partner biomolecule when they form the complex. By default this is described as an ambiguous distance restraints between all atoms of the source residue to all atoms of all target residue(s) that are assumed to be in close vicinity in the complex. The effective distance between all those atoms, $d_{iAB}^{eff}$ is calculated as follows:

$$d_{iAB}^{eff} = \left( \sum_{m_{iA}=1}^{N_{Aatom}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{Batom}} \frac{1}{d_{m_{iA}\,n_{kB}}^{6}} \right)^{-1/6} \quad \text{(Eq. 2.1)}$$

Here $N_{Aatom}$ indicates all atoms of the source residue on molecule A, $N_{resB}$ the residues defined to be at the interface of the target molecule B, and $N_{Batom}$ all atoms of a residue on molecule B. The $1/r^6$ summation somewhat mimics the attractive part of the Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the biomolecules are in contact. The AIRs are incorporated as an additional energy term to the energy function that is minimized during the docking. An upper distance limit

to the effective distance is enforced (typically 2 Å). If exceeded, the atom pairs that are part of the restraint experience an attractive force, otherwise the restraint is satisfied and the attractive force is zero. Since many atom-atom distances inversely contribute to the effective distance, an AIR is typically satisfied if any pair of atoms that are part of the restraint come within 3-5 Å of each other, depending on the degree of ambiguity (the total number of distances summed in Equation 2.1). As such the AIRs define a network of restraints between the possible interaction interface(s) of the molecules to be docked without defining their relative orientation minimizing the necessary search through conformation space needed to assemble the interfaces.

The ambiguous nature of these restraints easily allows experimental data that often provide evidence for the presence of a residue within an interface or residue-residue contacts to be used as driving force for the docking. AIRs are flexible in their use: they can also be setup for atom pairs, between residues selections or selected interfaces (useful for multi-body docking). This ensures that the full content of available information can be used for docking. The benefits of this are especially apparent for protein-DNA systems. The nucleotide units in a DNA helix are rather large in contrast to an amino acid. Their sugar phosphate backbone is most easily accessible at the surface of the helix while the bases are part of the "core" of the helix, accessible through the major and minor grooves. Various data sources provide information that is specific to parts of a nucleotide: ethylation interference for instance targets the phosphate group while mutation and conservation data targets the bases. For major groove-binding proteins the base restraints could even be targeted to specific atoms of the bases that face the major groove, providing a powerful way of limiting the search and ensuring correct positioning of the protein at the relevant interaction site.

**Implicit and explicit flexibility**

HADDOCK is able to deal with conformational changes in biomolecules upon complex formation. It accepts ensembles of starting structures as a means of implicit flexibility description and uses a semi-flexible refinement stage for explicit flexibility treatment. The latter is divided into two stages: side-chain and backbone optimization in torsion angle space and explicit solvent refinement in Cartesian space. Flexibility in the torsion angle space stage is by default automatically defined for those residues that are part of the interaction interface(s). For the protein(s) this approach is often sufficient, allowing side-chain and/or main-chain flexibility for those residues while the remainder of the protein is kept fixed. However, due to the dynamic behaviour of the DNA (as discussed before) its structure is likely to change conformation throughout the interface and, therefore, nearly the full structure should be defined flexible. Although it is possible to assign flexibility to all nucleotides of a DNA sequence, the approach is limited by the ability to maintain the correct helical conformation during the simulation. This problem can in part be accounted for by restraining the sugar-phosphate backbone dihedral angles, base planarity and Watson-Crick hydrogen bonds. Initial docking trials showed promising results reproducing the trends in the global conformational changes in terms of DNA bending and changes in the groove width. DNA helical deformations were nevertheless still a problem.

One of the benefits of the regular structure of the DNA polymer is the ease with which its 3D structure can be modelled. The conformation of a double stranded DNA structure can be defined in terms of 12 parameters (79): 6 parameters (Fig. 2.3a)

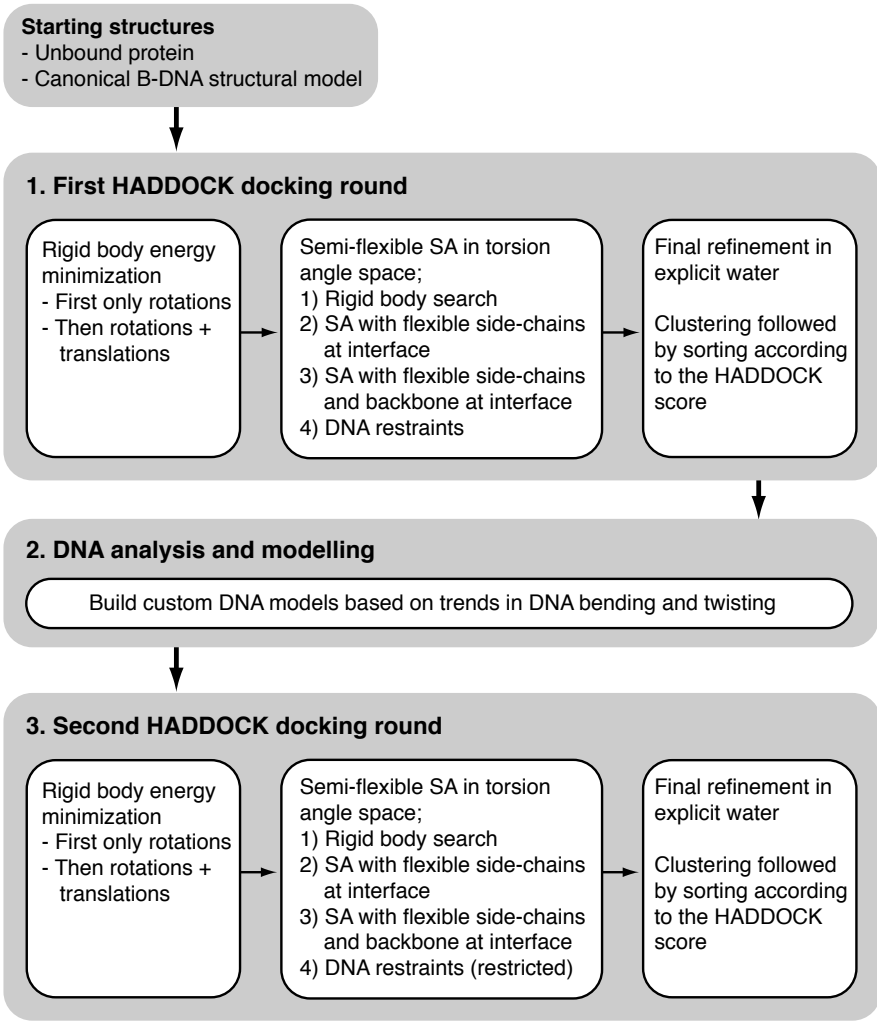





**Figure 2.4.** The two-stage protein-DNA docking approach using HADDOCK. **1**) Start with the unbound conformation of the protein(s) and a canonical B-DNA model, using experimental information to drive the docking and allowing for explicit flexibility for the amino acid residues at the interface and the full DNA. **2**) Analyze the DNA conformation in the docking results in search of trends in DNA bending and twisting and use this information to generate an ensemble of custom DNA models that sample the global conformation reflected by the observed trends. **3**) Use the new DNA ensemble as input for a second 'refinement' docking run during. SA; simulated annealing.

describe the orientation of two bases relative to each other and another 6 (Fig. 2.3b) describe the orientation of two successive base pairs (a base pair step) relative to each other. The geometrical relationships that have been found between the base pair step parameters and global DNA bending and twisting (46,82,238) allow for the custom modelling of double stranded DNA structures.

**A two-stage docking approach**

This opens new possibilities for dealing with DNA conformational changes in docking. Given that HADDOCK successfully introduces trends in DNA bending and twisting, this information can be used to generate an ensemble of custom DNA 3D structural models that sample the global conformation reflected by the observed trends. Such an ensemble should allow the sampling of a larger part of the relevant DNA conformational space than what is feasible within one round of semi-flexible refinement. This new DNA ensemble can subsequently be used for a second 'refinement' docking run during which the conformational freedom of the DNA in the semi-flexible refinement stage is restricted to prevent helical deformation.

In conclusion, the use of Ambiguous Interaction Restraints (AIRs) defined based on biochemical and/or biophysical data should be able to effectively reconstruct the correct interaction interface. The combination of explicit flexibility in HADDOCK with implicit flexibility by means of a DNA modelling stage should make this two-stage docking method (Fig. 2.4) better suited to deal with large DNA conformational changes. In chapter 3, this hypothesis is tested.

## Conclusions and Perspectives

A comprehensive understanding of interactions between biomolecules is of vital importance for the interpretation and prediction of biological processes in the living cell. Computational docking has proven to be a valuable addition to the tool chest of the structural biologist who studies these interactions. The docking field in general has seen major improvements over the past years. With an increased interested in processes implicated in gene duplication and regulation more effort is devoted to the development of effective protein-DNA docking methods. Over the last decade many novel approaches have been proposed to deal with the three main challenges that dominate the field: finding the proper interaction interface(s), dealing with conformational changes and effectively selecting the native complexes from the large number of generated solutions.

Protein-DNA docking evolved along the same lines as protein-protein docking. The early methods focused on identifying the proper interfaces on both the protein and DNA. For this, transcription factor - DNA test systems were considered in their bound state and docked in a rigid body fashion. These initial attempts were quite successful but their performance soon started to decline with the choice of different test systems and the consideration of unbound structures instead of bound ones. This decline was in part due to conformational changes upon complex formation and in part by the inability of scoring functions based on physicochemical properties to unambiguously identify the interaction interface. The latter limitation was improved with the use of additional information aimed at better identifying the interface(s) and the interactions made. The use of experimental or statistical information derived from the analysis of a large number of protein-DNA complexes is now common practice in many of the available docking methods. The consideration of conformational changes in protein-DNA systems, however, still poses a big challenge. The few methods that do account for them can handle protein side-chain rearrangements by explicit means but require implicit methods such as ensemble docking as an approximation of larger conformational changes. DNA conformational changes are almost exclusively dealt with by implicit means due to the inability of the various methods to maintain a proper helical conformation during the docking. As implicit flexibility is an approximation, there is often a

need for an additional refinement stage using molecular dynamics simulations to optimize the conformation of the protein and the DNA in the complex. Also recent promising developments follow this path of multi-stage docking. These include the use of normal mode analysis techniques implemented as an additional refinement stage or as a hinge detection method in cases where conformational changes can be described by rigid domain reorientations. However, disordered to ordered transitions and other large domain flexibility cannot be handled successfully by any of the methods to date. Moreover, none of the methods is able to dock more than two molecules simultaneously, which imposes problems due to the multi-component nature of many protein-DNA complexes. Furthermore, the developed methodology has often been validated using well-known transcription factor – DNA complexes in their monomeric or dimeric state. Considering the full spectrum of different protein-DNA complexes deposited in the RCSB protein databank, these complexes are certainly not the most challenging. A larger, more diverse set of test cases would be a welcome addition for method development and validation.

A proper handling of conformational changes is vital to the prediction of protein-DNA complexes. It is expected that the problem of predicting and describing conformational changes will dominate the docking field for quite some time.

# Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility

Marc van Dijk,
Aalt D.J. van Dijk,
Victor Hsu,
Rolf Boelens,
Alexandre M.J.J. Bonvin

*Nucleic Acids Research*
(2006), **34**: 3317-3325

Chapter 3

Intrinsic flexibility of DNA has hampered the development of efficient protein-DNA docking methods. In this study we extend HADDOCK (High Ambiguity Driven DOCKing (86)) to explicitly deal with DNA flexibility. HADDOCK uses non-structural experimental data to drive the docking during a rigid body energy minimization, and semi-flexible and water refinement stages. The latter allow for flexibility of all DNA nucleotides and the residues of the protein at the predicted interface. We evaluated our approach on the monomeric repressor-DNA complexes formed by bacteriophage 434 Cro, the Escherichia coli Lac headpiece, and bacteriophage P22 Arc. Starting from unbound proteins and canonical B-DNA we correctly predict the correct spatial disposition of the complexes and the specific conformation of the DNA in the published complexes. This information is subsequently used to generate a library of pre-bent and twisted DNA structures that served as input for a second docking round. The resulting top ranking solutions exhibit high similarity to the published complexes in terms of RMS deviations, intermolecular contacts and DNA conformation. Our two-stage docking method is thus able to successfully predict protein-DNA complexes from unbound constituents using non-structural experimental data to drive the docking.



## Introduction

Computational docking has proven to be a valuable tool in the study of biomolecular complexes (120,281). In particular, the field of ab initio protein-protein docking has made considerable progress as illustrated by recent results from the community-wide CAPRI experiment (Critical Assessment of Predicted Interactions (138,218)). However, where this field has in many ways matured, the development of docking methods to model protein-DNA interactions has lagged behind. These play an important role in recognition and gene expression (260). Powerful protein-DNA docking methods would thus be of great benefit for their study. However, two particular problems have hampered the development of efficient docking methods: the sparsity of the information to define the DNA binding interface and the inherent flexibility of DNA. For protein-protein docking there is often enough information available (e.g. from sequence, conservation or biological knowledge) to identify the interaction surfaces of the docking partners. This information can be used to drive the docking (323) and limit the conformational space to be searched. Identification of the interaction surface on DNA is less straightforward

than on proteins: There is still no general recognition code and the global conformation of the DNA can play an important role in modulating the eventual interaction surface (244). DNA indeed often exhibits large conformational changes upon binding to a protein, which can greatly alter the shape of the interaction surface. Due to this, the total conformational space that needs to be searched in order to find favourable conformations becomes even larger. Flexibility in DNA can be separated into global and local components (315). Global flexibility is constrained to two primary motions: bending and twisting. It results from a combination of conformational changes in the flexible base pairs and sugar-phosphate backbone. Allowing for global and local flexibility in DNA during docking while maintaining the relevant conformation is a major challenge in protein-DNA docking.

In the last few years several methods have been developed to solve one or both of these problems, each with varying degrees of success. The program FTDOCK (8) has been used to perform a large search through conformational space by rotating and translating the protein along the DNA while evaluating shape and electrostatic complementarity; an approximation of flexibility was achieved by allowing some degree of overlap between protein and DNA in the scoring. In another approach, a library of pre-bent DNA structures was used to minimize the search through DNA conformational space (270); a selection was made based on structures that could be electrostatically preorientated in the potential of the protein and these were rotated and translated with respect to the protein. To account for some degree of local flexibility protein side-chains and DNA base pairs were allowed to move in two separate refinement stages. Knegtel *et al.* developed MONTY (156) which uses a Monte Carlo search allowing for flexibility in both protein and DNA and experimentally determined contacts to drive the docking. The initial position of the protein in the predicted complex should, however, not deviate too much from that of the actual complex; small deviations in the position of the protein with respect to the interaction interface of the DNA resulted in DNA curling around the protein. Tzou *et al.* (317) modelled the CAP-DNA and Rep-DNA systems from the repressors in their bound conformation and canonical B-DNA in a series of molecular mechanics and dynamics simulations using distance restraints derived from a statistical analysis of homologous protein-DNA complexes. This method successfully introduced DNA bending and local opening of the major groove. All of these docking procedures were able to make predictions that were representative of the published complexes in terms of spatial disposition. Only a few methods allowed for flexibility of the DNA and protein side-chains during the docking. They, however, required extensive knowledge to position the two components relative to each other (317) and problems were encountered in the absence of such information (156).

Here we demonstrate that both global and local DNA flexibility can successfully be accounted for in protein-DNA modelling using HADDOCK (86), a computational docking approach developed in our group. HADDOCK makes use of available experimental and bioinformatics data to drive the docking process (323). Its successful use in NMR-based structure calculations of protein-DNA and protein-RNA complexes has been shown before (148,149,159,333). Global and local DNA flexibility is introduced in the docking by allowing the DNA sugar-phosphate backbone and DNA base pairs to sample conformations during a semi-flexible refinement stage and by starting the docking from a library of pre-generated DNA structures representing various degrees of conformational flexibility. The latter allows for the sampling of a larger conformational

space. Flexibility in the protein is introduced as described previously (86), first along the side-chains at the interface and then for both backbone and side-chains. We demonstrate here the feasibility of this approach for three repressor complexes in their monomeric form: Cro from bacteriophage 434 (222), the *Lac* headpiece of *E. coli* (297) and Arc from bacteriophage P22 (257). The first two recognize the DNA major groove via a α-helix/turn/α-helix motif and the last one via a two-stranded antiparallel β-sheet motif. To drive the docking we make use of mutation data, sequence/structure conservation, DNA footprinting and ethylation interference data. We show that our approach is successful in predicting protein-DNA complexes from unbound constituents by accounting for both global and local DNA flexibility during the docking.

## Results

### Bound rigid body docking

The use of readily available biochemical and/or biophysical information can alleviate the lack of a general recognition code for protein-DNA interactions. HADDOCK uses this information encoded as Ambiguous Interaction Restraints (86) to drive the docking; this reduces the necessary search through interaction space and increases the fraction of unique solutions. In the definition of AIRs we distinguish between active and passive residues. Active residues are defined as those important for the interaction based on conservation (HSSP, (269)), mutation or ethylation interference data or any other appropriate experimental data. Passive residues are defined as the solvent-accessible neighbours of active residues (Table 3.3).

**Table 3.1.** RMS deviations from the target and fraction of native contacts for the top five ranking docking solutions of the best cluster.

| | RMSD (Å) | | | | Fnat[e] |
|---|---|---|---|---|---|
| | total[a] | Interface[b] | Backbone[c] | DNA[d] | |
| **Cro – O1R** | | | | | |
| bound | $0.27_{0.00}$ | $0.24_{0.00}$ | $0.28_{0.00}$ | $0.00_{0.00}$ | $0.88_{0.00}$ |
| unbound rigid | $2.62_{0.01}$ | $2.37_{0.06}$ | $1.92_{0.02}$ | $2.31_{0.00}$ | $0.53_{0.12}$ |
| unbound flex. | $2.30_{0.07}$ | $2.07_{0.12}$ | $1.80_{0.09}$ | $1.97_{0.15}$ | $0.80_{0.07}$ |
| DNA lib. | $1.99_{0.05}$ | $1.69_{0.06}$ | $1.51_{0.09}$ | $1.46_{0.07}$ | $0.94_{0.00}$ |
| **Lac – O1** | | | | | |
| bound | $0.34_{0.00}$ | $0.31_{0.00}$ | $0.36_{0.00}$ | $0.00_{0.00}$ | $0.89_{0.00}$ |
| unbound rigid | $2.84_{0.00}$ | $2.88_{0.00}$ | $2.56_{0.00}$ | $1.71_{0.00}$ | $0.33_{0.00}$ |
| unbound flex. | $2.64_{0.10}$ | $2.56_{0.12}$ | $2.41_{0.12}$ | $1.90_{0.18}$ | $0.51_{0.03}$ |
| DNA lib. | $2.33_{0.06}$ | $2.29_{0.08}$ | $2.06_{0.08}$ | $1.57_{0.09}$ | $0.54_{0.01}$ |
| **Arc – operator** | | | | | |
| bound | $0.22_{0.00}$ | $0.23_{0.00}$ | $0.19_{0.00}$ | $0.00_{0.00}$ | $0.95_{0.00}$ |
| unbound rigid | $2.58_{0.01}$ | $2.58_{0.01}$ | $1.97_{0.02}$ | $2.52_{0.00}$ | $0.43_{0.00}$ |
| unbound flex. | $2.24_{0.08}$ | $2.13_{0.10}$ | $1.64_{0.10}$ | $1.88_{0.15}$ | $0.50_{0.04}$ |
| DNA lib. | $2.20_{0.15}$ | $2.19_{0.19}$ | $1.73_{0.15}$ | $1.99_{0.11}$ | $0.51_{0.08}$ |

Average RMSD values (Å, standard deviation in subscript) calculated over the entire complex (a), the interface (b), the backbone (c) and the DNA (d) for the five top ranking solutions. The RMSDs are reported for the bound rigid-body docking (bound), unbound docking before (unbound rigid) and after semi-flexible refinement (unbound flex.) starting from canonical B-DNA, and unbound semi-flexible docking using a library of pre-bent and twisted DNA as input structures (DNA lib.). (e) Fnat is the fraction of native contacts.

We first evaluated the use of AIRs in protein-DNA docking for the three selected complexes by bound docking (i.e. the reconstruction of the complexes from their separate components). Since the molecules are already in their bound conformation no flexible segments were defined and only rigid-body docking was performed. The best docking solutions for each of the Lac, Arc and Cro repressor in complex with their operators exhibit high similarity with the published complexes based on RMS deviations and intermolecular contacts (Table 3.1); all base-specific intermolecular contacts are recovered.

In the biologically relevant complexes the repressors are bound as dimers that are symmetrically oriented on the two recognition sites of the operator. In this study we use the repressors in their monomeric form (in this form the Arc repressor is a symmetrical dimer). Symmetry in the AIR set and in the shape of the protein-DNA interaction surface can result in false positives: these are structures with a favourable HADDOCK score (weighted sum of several energy terms, see Materials and Methods) but with one of the two components 180° rotated with respect to the published complex. To minimize the occurrence of false positives 180° rotated solutions were systematically sampled during the rigid body docking stage. For this, a 180° rotation around a vector defined by the centres of masses of the interfaces of the protein and DNA was applied and the resulting conformation again minimized. The solution with the lowest HADDOCK score was kept. Using this approach the amount of false positives after the rigid body docking stage was reduced from ~70% to ~40%. In subsequent unbound docking runs including flexibility we selected the best 20% of all solutions from the rigid body docking stage based on their HADDOCK score. Due to the sampling of 180° rotations this subset contained no false positives for the Cro and Lac repressor/operator complexes (Fig. 3.1). Because of the intrinsic symmetry of the

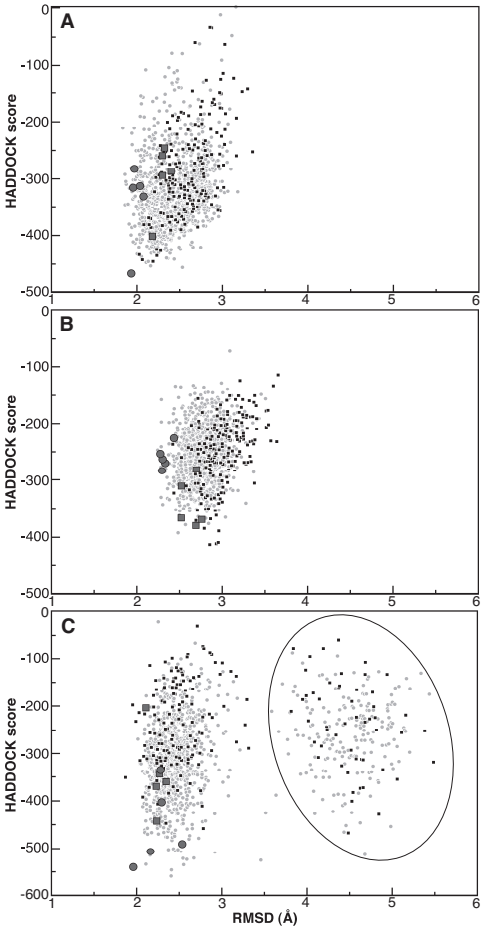



***Figure 3.1.** HADDOCK score versus RMSD from the target (all heavy atoms of the complex) for the Cro (A), Lac (B) and Arc (C) repressors in complex with their operator. Solutions of the unbound flexible docking with canonical B-DNA are shown as small black squares with the five top ranking solutions identified by red squares. Solutions from the docking using a library of pre-bent and twisted DNA structures are shown as small orange circles with the top five ranking solutions identified by red circles. False positives for Arc are shown within a solid ellipse: These correspond to solutions in which the repressor is shifted one or two base pairs along the DNA.

Arc repressor, 180° rotated symmetrical solutions are similar and can thus not be distinguished. Therefore the problem of rotational false positives does not apply to the Arc repressor. In unbound docking false positives were obtained that correspond to shifted false positives. These are solutions in which the repressor is shifted one or two base pairs up or downstream of the true interaction surface on the DNA (Fig. 3.1).

**Unbound semi-flexible docking to B-form DNA**

We used the AIR sets to dock an ensemble of NMR structures of the unbound repressors to canonical B-DNA (chosen for it's biological relevance). In contrast to the previous bound docking runs in which only rigid body docking was performed, we now included flexibility in a semi-flexible refinement stage: side-chains and backbone of the protein at the predicted interface and the entire DNA were allowed to sample additional conformations. A set of restraints was imposed on the DNA that allowed for local flexibility but preserved the overall helical conformation (see Materials and Methods). The final refined structures were clustered based on their pair wise RMSD matrix. The best cluster was selected based on the HADDOCK score.

The solutions in the selected clusters appeared to be very similar with respect to the protein and the spatial disposition of the complex but less similar on the level of the DNA conformation. An analysis of the base-pair and base-pair step parameters of the DNA in the selected clusters revealed a higher variation in buckle, propeller, roll and tilt than in other parameters (Table 3.2). Previous studies have also observed a larger variation for these parameters in both

**Table 3.2.** Average DNA base-pair and base-pair step parameters.

| | Cro | | | Lac | | | Arc | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Ref.** | **Docking from:** | | **Ref.** | **Docking from:** | | **Ref.** | **Docking from:** | |
| **Parameters** | 3cro | bDNA | Lib. | 1lcc | bDNA | Lib. | 1bdt | bDNA | Lib. |
| Twist ($35.9°_{0.9}$) | $34.4_{5.3}$ | $36.4_{1.0}$ | $34.9_{3.5}$ | $34.2_{5.4}$ | $36.8_{0.7}$ | $36.5_{3.6}$ | $32.5_{4.1}$ | $34.6_{1.2}$ | $35.5_{3.3}$ |
| Roll ($-0.2°_{2.3}$) | $2.5_{3.2}$ | $-0.2_{2.0}$ | $1.0_{8.1}$ | $2.6_{11.2}$ | $0.3_{1.7}$ | $0.2_{10.4}$ | $3.3_{5.5}$ | $4.2_{1.9}$ | $1.0_{7.7}$ |
| Tilt ($0.0°_{0.1}$) | $0.5_{3.7}$ | $0.0_{2.0}$ | $0.4_{5.4}$ | $-2.7_{7.9}$ | $0.2_{1.6}$ | $0.2_{4.9}$ | $-0.3_{3.3}$ | $0.9_{1.5}$ | $0.4_{5.9}$ |
| Rise ($3.4_{0.0}$ Å) | $3.4_{0.3}$ | $3.3_{0.2}$ | $3.4_{0.4}$ | $3.2_{0.2}$ | $3.3_{0.2}$ | $3.3_{0.3}$ | $3.3_{0.2}$ | $3.3_{0.1}$ | $3.3_{0.4}$ |
| Slide ($0.3_{0.2}$ Å) | $-0.4_{0.4}$ | $0.0_{0.1}$ | $-0.6_{0.6}$ | $-0.4_{0.7}$ | $0.2_{0.2}$ | $0.1_{0.7}$ | $-0.4_{0.7}$ | $0.0_{1.7}$ | $-0.3_{0.5}$ |
| Shift ($0.0_{0.1}$ Å) | $0.0_{0.5}$ | $0.1_{0.1}$ | $0.0_{0.6}$ | $-0.1_{0.7}$ | $0.1_{0.3}$ | $0.0_{0.5}$ | $-0.1_{0.9}$ | $0.1_{0.3}$ | $0.1_{0.8}$ |
| Opening ($-3.3_{2.5}$ Å) | $-4.5_{4.8}$ | $-4.6_{2.2}$ | $-3.3_{4.0}$ | $-6.7_{7.9}$ | $-2.0_{2.8}$ | $-2.0_{3.8}$ | $0.4_{4.3}$ | $-0.8_{2.0}$ | $-0.8_{4.7}$ |
| Propeller ($-10.2°_{7.3}$) | $-14_{5}$ | $-7.5_{4.4}$ | $-0.9_{12.7}$ | $-14.6_{4.4}$ | $-8.5_{5.0}$ | $-9.3_{10.1}$ | $-4.3_{8.3}$ | $-4.7_{3.8}$ | $-1.1_{14.1}$ |
| Buckle ($0.1°_{0.1}$) | $1.0_{8.1}$ | $-1.4_{5.1}$ | $-0.6_{10.8}$ | $-6.9_{13.2}$ | $4.6_{3.5}$ | $-0.2_{11.8}$ | $-2.7_{6.5}$ | $5.2_{4.9}$ | $-2.5_{13.5}$ |
| Stagger ($0.1_{0.0}$ Å) | $-0.1_{0.5}$ | $-0.1_{0.2}$ | $-0.3_{0.6}$ | $0.1_{0.8}$ | $-0.1_{0.2}$ | $0.1_{0.5}$ | $0.0_{0.3}$ | $-0.2_{0.3}$ | $-0.2_{0.5}$ |
| Stretch ($-0.1_{0.0}$ Å) | $-0.3_{0.2}$ | $-0.1_{0.1}$ | $-0.2_{0.1}$ | $-0.1_{0.2}$ | $-0.2_{0.1}$ | $-0.1_{0.1}$ | $-0.2_{0.1}$ | $-0.1_{0.1}$ | $-0.1_{0.1}$ |
| Shear ($0.0_{0.1}$ Å) | $0.2_{0.5}$ | $0.1_{0.0}$ | $0.0_{0.3}$ | $-0.3_{0.5}$ | $0.1_{0.1}$ | $-0.1_{0.2}$ | $-0.1_{0.3}$ | $0.0_{0.4}$ | $-0.1_{0.2}$ |
| **Correlations** | | | | | | | | | |
| Roll-twist (0.26) | -0.47 | -0.55 | -0.44 | -0.65 | -0.61 | -0.76 | -0.85 | -0.16 | -0.23 |
| Roll-slide (0.30) | -0.40 | -0.43 | -0.37 | -0.65 | -0.48 | -0.61 | -0.44 | 0.00 | -0.43 |

Average parameters with standard deviations in subscript are shown for the published complexes (Ref.) and the top five ranking solutions from unbound flexible docking starting from canonical B-DNA (bDNA) and from a library of pre-bent and twisted DNA as input structures (Lib.). For comparison, the average values for the canonical B-DNA input structure are shown in the left column between brackets next to each parameter.

free and bound DNA when it is bending and twisting (81,237,260,305,315). This is not surprising as buckle, propeller, roll and tilt parameters are less restricted by Watson-Crick hydrogen bonds and the conformation of the sugar-phosphate backbone, than is the case with the other parameters. However, their large variation occasionally resulted in an overall loss of B-DNA conformation in the docking solutions as assessed by 3DNA (194). These solutions, however, did not have worse HADDOCK scores than solutions with a smaller variation in the noted parameters. They could, however, in most cases be distinguished by their higher DNA deformation energy. For this we calculated the combined base pair and base-pair step deformation energy for every solution in the selected cluster and ranked them according to this energy term (see Materials and Methods); The ranked solutions were checked on having a general B-DNA conformation and the best 5 were selected. This procedure proved successful in selecting solutions that are in better agreement to the published complexes in terms of RMSD values (Fig. 3.1).

To assess the effect of flexibility on the docking we compared the top ranking solutions after the semi-flexible refinement stage with their initial conformation after rigid body docking: the results show a clear improvement in RMSD from the published structure of the complex and fraction of native contacts (Table 3.1). Analysis of the DNA revealed that the backbone torsion angles were all located in the most populated regions as derived from a statistical analysis of non-complexed DNA structures (29,279) (data not shown). Base-pair buckle, propeller, tilt and roll parameters, which are at the origin of overall DNA bending and twisting, showed larger differences between the published complexes and the rigid body docking solutions than after introduction of flexibility (Table 3.2). Base-pair opening, stagger, stretch and shear parameters and base-pair step twist, slide and shift parameters showed little differences. In all three complexes the DNA is slightly bent towards the protein. In this respect tilt rotation is reported to be both statistically and energetically less favourable than roll rotation (82,93,117,359). This relationship is observed in the published complexes and the top ranking docking solutions as they show smaller variations in tilt than in roll. Statistical analysis of crystal structures has revealed that a positive change in roll is often accompanied with unwinding and negative slide (93,116,308). In our best solutions we also witness that roll is negatively correlated with both twist and slide (Table 3.2); more precisely, twist values below 36° are often accompanied with negative sliding in bent DNA. This relation is observed at the interface of the top ranking docking solutions (the central four base-pair steps in panels D,E,F,G,H and I of Fig. 3.2). On a global level the distribution of major groove widths over the different base-pair steps followed a trend similar to the published complexes (Fig. 3.2 A,B,C).

**Unbound docking from custom-build DNA libraries**

The results above show that the introduction of flexibility results in the prediction of a more native-like complex in comparison to rigid body docking. To account for even larger DNA conformational changes we explored the possibility of using a library of pre-bent and twisted DNA structures as input structures for the docking procedure. Although the DNA in the best clusters of the flexible docking runs starting from canonical B-DNA showed variation on a local level (e.g. buckle, propeller, roll and tilt parameters) the global conformation of all solutions was quite similar. Analysis of the resulting DNA conformations provided information in the form of bend-angles and the width of the major groove, which was used to construct custom DNA libraries. For

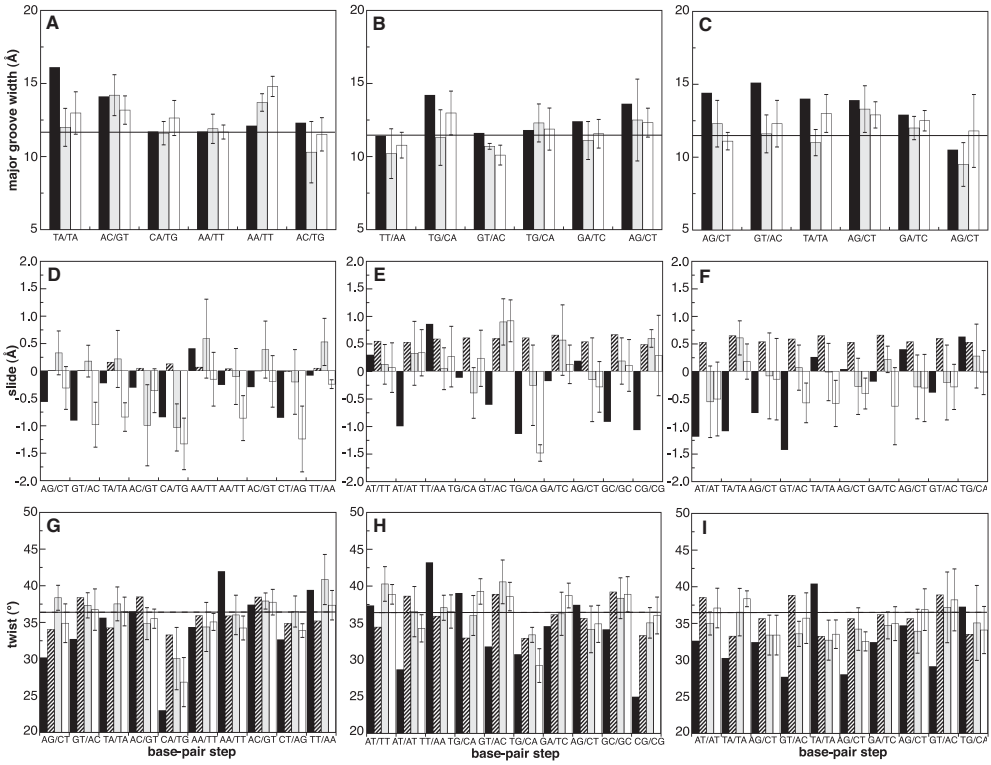



**Figure 3.2.** Major groove width, slide and twist parameters of the five top ranking solutions of the Cro (A,D,G), Lac (B,E,H) and Arc (C,F,I) repressor/operator complexes. Average values plus standard deviations for the solutions of the unbound flexible docking with canonical B-DNA are shown as grey bars and those using a library of pre-bent and twisted DNA structures are shown as white bars. The values as measured in the published complexes are presented as black bars and those of the canonical B-DNA input structures as striped bars for slide (D,E,F) and twist (G,H,I) and as a horizontal solid line for the major groove width. All values for the major groove width are corrected by 5.8 Å to account for van der Waals radii of the phosphate groups. A value of 36° twist is presented as a dashed line for clarification of the twist-slide relationship.



the Cro/O1R complex the major groove width increased from 11.6 Å (canonical B-DNA) to 12.5 ± 0.5 Å and the DNA adopted a curve towards the protein of 9.4 ± 3.6°. For the Lac and Arc repressors in complex with their operator similar events occur, resulting in major groove widths of 11.3 ± 0.4 Å and 12.2 ± 0.8 Å and curves towards the protein of 11.3 ± 3.8° and 12.9 ± 5.2°, respectively. Based on this information we constructed for each operator five DNA structures that sample values around the averaged major grooves widths and bend angles from the previous docking runs. Docking from these libraries using the flexible protocol described above resulted in solutions with twist and slide parameters as well as major groove widths in better agreement with those of the published complexes (Fig 3.2). The overall results (Table 3.1) demonstrate that the use of a custom library of pre-bent and twisted structures improves the prediction structures of the complexes as assessed by RMSD values, intermolecular contacts and DNA conformation. Only for the Arc repressor/operator complex did the

use of a custom DNA-library not result in a significant improvement compared to canonical B-DNA docking. The best docking solutions superimposed onto their reference structure are presented in Figure 3.3.

## Discussion

Our modelling of protein-DNA complexes is based on Ambiguous Interaction Restraints to drive the docking process. These are essential in successfully positioning the protein at the interface of the DNA and, together with flexibility, influence DNA bending in the semi-flexible refinement stage. We used a limited number of easily obtainable experimental data to define the restraints. These were nevertheless sufficient to accurately predict the conformation of the DNA in the complex when starting from canonical B-DNA. This information subsequently allowed us to refine our models by performing docking from a custom-built DNA library instead of canonical B-DNA. This two-stage docking approach significantly improved the conformation of the DNA compared to canonical B-DNA and the overall docking results. The conformation of the protein did not differ much from the starting structures.

In this study we did not investigate the effects of a variable number or type of restraints on the docking results. From analogous protein-protein docking studies it is known that the amount of- or the ambiguity in the data can influence the reproducibility of the docking. HADDOCK allows the random deletion of a fraction of the restraints for each docking trial to account for errors in their definition, an approach that has proved successful in the past (324). This option was not used in this study. The AIR restraints were defined with an upper distance limit of 2.0 Å that can affect the packing of the docking solutions. For the Lac/O1 and Arc/operator complexes the buried surface area was comparable to that of the reference (1496 ± 103 Å versus 1560 Å and 1990 ± 155 Å versus 2072 Å respectively). For the Cro/O1R complex the buried surface area of the top ranking solutions was larger than that of the reference (1694 ± 52 Å versus 1453 Å). The tighter packing might contribute to the significant increase in the fraction of native contacts (Table 3.1) for the Cro/O1R complex, with respect to the other two test systems.

We have demonstrated that the use of readily available non-structural experimental data and the incorporation of DNA flexibility during the docking significantly improve



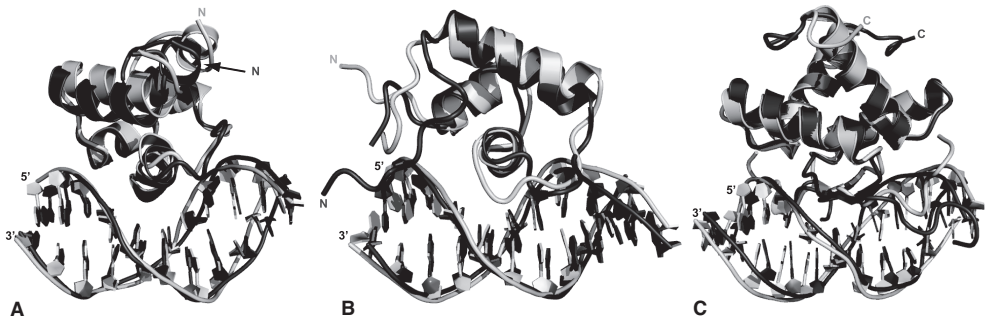



***Figure 3.3.** Best solutions of the unbound flexible docking using a library of pre-bent and twisted DNA structures (blue) superimposed on the reference structure (yellow): Cro-O1R (**A**), Lac-O1 (**B**), Arc-operator (**C**). The structures were superimposed on all heavy atoms of the interface residues (Interface RMSD values: Cro, 1.62 Å; Lac, 2.02 Å; Arc, 1.90 Å). The figures were generated using Pymol (DeLano Scientific LLC, www.pymol.org).

repressor-DNA complex prediction in comparison to rigid-body docking. The method successfully predicted global conformational changes taking place in the DNA upon complexation. The information extracted from these results is sufficient to refine the models by starting a second docking round from custom-built DNA-libraries of pre-bent and twisted structures.
The flexible protein-DNA docking approach described in this paper has biological implications since it can benefit studies of protein-DNA interactions at several levels. It can be used to generate models of protein-DNA complexes when the structure of the unbound protein is known and suitable experimental data is available. It is also applicable to study the effects of mutations or different operator sequences on complex formation. In addition, it can assist in experimental structural studies: it can for example speed up structure determination of protein-DNA complexes by NMR by providing initial models to guide the tedious NMR analysis and assignment process. In summery, by allowing the inclusion of a large variety of experimental and/or bioinformatics data, together with a flexible description of the DNA, the proposed docking approach should be a useful tool in structural studies of protein-DNA and even protein-RNA interactions provided suitable RNA models are available for the latter.



## Materials and methods

### Initial structures for protein and DNA

The coordinate files of all proteins and protein-DNA complexes were obtained from the RCSB Protein Data Bank (30). The PDB entry codes of the respective complexes and their unbound components are as follows: 3CRO, crystal structures of the Cro/O1R complex (222); 1ZUG, NMR ensemble of the unbound Cro monomer (245); 1LCC, NMR ensemble of the Lac/O1 complex (60); 1LQC, NMR ensemble of the unbound monomer of the Lac-headpiece (295); 1BDT, crystal structure of the Arc/DNA complex (276) and 1ARQ, NMR ensemble of the unbound Arc monomer (37). The monomeric reference structures for Cro and Arc (right halfside) were extracted from the dimeric PDB structures.
Models of canonical B-DNA were constructed with the nucleic acid analysis and rebuilding program 3DNA (194), using the fiber models provided by Arnott *et al.* (50). All hydrogens were added according to the standard assigning scheme of CNS followed by a short energy minimization step during the initiation stage in HADDOCK. Base-pair and base-pair step parameters of the resulting type BII B-DNA starting structures are shown in Table 3.2. The DNA backbone torsion angles are: $\alpha = 309°$, $\beta = 159°$, $\gamma = 37°$, $\delta = 146°$, $\varepsilon = 218°$, $\zeta = 191°$, $\chi = 260°$ and the sugar pseudo-rotation phase angle (P) = 155°, the sugar pucker was thus in the C2'-endo conformation.
Custom DNA libraries for the three operator sequences were generated by manipulation of the base-pair step parameters of their respective B-DNA structures using 3DNA. The introduction of curvature was accomplished by changing the value of roll using the equation (46):

$$Rn = k \cos (T * n\theta) \qquad \text{(Eq. 3.1)}$$

Where $R_n$ is the roll value for each base-pair step in one helical turn, n is the number of base-pair steps in one helical turn, k is the average curvature for each base-pair step in one helical turn and T is the value for twist. The direction of the curvature in Cartesian space can be controlled by changing the phase ($\theta$) of the cosine function. The positive linear relationship between the value of the slide parameter and the width of the major groove was used to adjust the major groove width.

### Restraints used in the docking

*Ambiguous Interaction Restraints (AIR, Table 3.3)*: All active residues have a relative solvent accessibility higher than 50% as calculated

**Table 3.3.** Definition of the AIRs for the three repressor/operator systems.

| | protein | DNA | reference |
|---|---|---|---|
| **Cro – $O_1R$** | | | |
| Active | K27[a],Q29[a],S30[a],L33[ac] | T3[b],A4[ab],C5[a],A6[a],G30[b],T31[abc],T32[abc],T33[a],G 34[a],T35[a] | (42,125,164,345) |
| Passive | R10,K27,Q29,K40,R41,P42 | - | |
| **Lac – O1** | | | |
| Active | T5[ac],S16[a],Y17[c],Q18[c],R22[c],V30[c] | T4[b],G5[ab],T6[a],G7[a],A8[a],C14[b],T15[ab],C16[a],A17[a], C18[a] | (19,97,154,211) |
| Passive | H29,S31 | - | |
| **Arc – operon** | | | |
| Active | F10[a],R13[a],S32[a] | T1[b],A2[b],T3[b],G5[b],T6[a],A7[a],G8[a],A9[a],A14[b],C15[b],T 16[b],C17[a],T18[a],A19[a] | (330) |
| passive | Q9,N11,R16,D20,R23 | - | |

The Arc monomer is composed of two symmetric subunits and only the restraints for one subunit are shown. [a]Conserved residues derived from the database of homology-derived secondary structure of proteins (HSSP); [b]Ethylation interference; [c]Mutagenesis data.

with NACCESS (134). Residues located in the predicted interaction interface or in a continuous stretch of residues near the predicted interaction interface for which no information is available were defined as passive. AIR restraints for the protein were defined based on sequence conservation (HSSP, (269)) and mutation data. For the DNA only active residues were defined. The recognition sequences of the operators have been determined using DNA footprinting methods before the experimental structures of the actual complexes became available. This information was used in our docking procedure to define interaction restraints. For those bases shown to be involved in specific interactions with the repressor, only atoms able to interact by hydrogen-bond or non-bonded interactions were defined. Based on ethylation interference experiments, only the oxygen atoms of phosphate groups shown to interact with the repressor were defined as active.

*DNA Restraints:* In order to preserve the helical conformation of DNA the following restraints were defined: planarity restraints for the purine and pyrimidine rings were introduced, and the sugar pucker was restrained to the C2'-endo conformation. Watson-Crick base-pairs were defined and hydrogen bond lengths of the input structure (either the starting DNA conformation or the conformation obtained after semi-flexible refinement prior to water refinement) were measured and restricted to ± 0.05 Å. In a similar way the dihedral angles of the sugar-phosphate backbone of the input structure (imp) were measured and used as restraints. (Restricted to: $\alpha = \alpha_{imp} \pm 10°$, $\beta = \beta_{imp} \pm 40°$, $\gamma = \gamma_{imp} \pm 20°$, $\delta = \delta_{imp} \pm 50°$, $\varepsilon = \varepsilon_{imp} \pm 10°$ and $\zeta = \zeta_{imp} \pm 50°$).

**Docking protocol**

Our docking protocol consists of i) rigid-body docking, ii) semi-flexible refinement stage and iii) final refinement in explicit solvent.

*i) Rigid-body docking:* A total of 100 structures were generated for each protein-DNA combination from the ensembles of starting structures. Each docking attempt was performed 10 times and the solution with the lowest HADDOCK score was kept. For each protein we used an ensemble of 10 NMR structures; thus 1000 rigid body docking solutions were generated for each of the three canonical B-DNA docking runs and 5000 structures were generated for each of

the DNA library docking runs (5 different pre-bent and twisted DNA structures and 10 protein structures resulting in 50 different combinations). For the docking of the protein and DNA in their bound conformation a total of 1000 structures were generated. Systematic sampling of 180° rotated solutions was used in the rigid body docking stage to minimize the occurrence of false positives (principles described in the Results section). This basically doubled the number of docking trials bringing the total to 20000 and 100000 evaluations for docking from canonical B-DNA and DNA libraries respectively.

*ii) Semi-flexible refinement:* Of all structures generated in the rigid body docking stage the best 20% based on the HADDOCK score were further refined in the semi-flexible refinement stage consisting of three parts: rigid-body Torsion Angle Dynamics (500 MD steps at 2000 K and 500 MD cooling steps to 500 K with a 8 fs time step), semi-flexible simulated annealing stage (1000 MD steps from 1000 K to 50 K with 4 fs time steps) with the side-chains of the protein residues at the interface and the complete DNA (excluding terminal base-pairs) allowed to move and a final semi-flexible simulated annealing stage (1000 MD steps from 300 K to 50 K with 2 fs time steps) with both side chains and backbone of the protein residues at the interface and the complete DNA (excluding terminal base-pairs) allowed to move.

*iii) Water refinement:* This final stage consists of a gentle refinement (100 MD heating steps at 100, 200 and 300 K followed by 750 sampling steps at 300 K and 500 MD cooling steps at 300, 200 and 100 K all with 2 fs time steps) in a 8 Å shell of TIP3P water molecules (145).

Semi-flexible segments for the proteins were defined as: residues 7-20, 24-37 for Cro, residues 6-30, 50-56 for Lac and residues 1-17, 54-70 for Arc. In all cases the complete DNA, excluding the terminal base-pairs, were defined as semi-flexible.

**Scoring**

A HADDOCK score is defined to rank the structures after each docking stage. It is a weighted sum of intermolecular electrostatic (Elec), van der Waals (vdW), desolvation (Dsolv) and AIR energies, and a buried surface area (BSA) term: rigid-body score = 1.0*Elec + 1.0*vdW – 0.05*BSA + 1.0*Dsolv + 1.0*AIR, final score = 1.0*Elec + 1.0*vdW + 1.0*Dsolv + 1.0*AIR. A cluster analysis was performed on the final docking solutions using a minimum cluster size of 4. The cut off for clustering was manually determined for each docking run. The RMSD matrix was calculated over the backbone atoms of the interface residues of the DNA after fitting on the interface residues of the protein. Final structures within a cluster were selected according to their summed base pair and base-pair step deformation energies and the conformation of the helix (classified as B-DNA). Deformation energies were calculated with an extension script of 3DNA (provided by Marc Parisien, University of Montreal, Canada) using the statistical population preferences as determined by Olson *et al* (237) and Lankas *et al* (173).

Default HADDOCK (version 2.0_devel) parameters were used except for the dielectric constant (epsilon) that was set to 78 for the vacuum part of the protocol. To speed up calculations, non-polar hydrogens were omitted. Inter- and intramolecular energies were evaluated using full electrostatic and Van der Waals energy terms with a 8.5 Å distance cut-off. OPLSX non-bonded parameters from the parallhdg5.3.pro parameter file (188) were used for the protein. Topology and linkage parameter files for the DNA were taken from the CNS (41) distribution (dna-rna-allatom.top and dna-rna-allatom.param respectively). The HADDOCK package is freely available to academic users (see http://www.nmr.chem.uu.nl/haddock).

**Analysis**

RMSD values of the complexes were calculated using ProFit (Martin, A.C.R., www.

bioinf.org.uk/software/profit). All heavy atoms were used to calculate the RMSD of the total complex, of the DNA and of the interface. The interface was composed of residues: 15-44/3-7, 31-37 of Cro/O1R: 6-32/4-10, 13-19 of Lac/O1 and 8-36, 61-89/1-9, 13-21 of Arc/repressor. The backbone RMSD was calculated using all P and Cα atoms of the complex. Residues in the flexible termini of the protein (having either high B-factors in the X-ray structures or poorly defined in the NMR reference structures) were left out of the calculation. Intermolecular contacts were evaluated using LIGPLOT (336) using a 5Å cut-off. The fraction of native contacts (Fnat) is defined as the number of native intermolecular contacts on a nucleotide-residue basis (hydrogen-bonded and non-bonded) identified in a docking solution divided by the total number of contacts in the reference structure. Values for base-pair and base-pair step parameters as well as torsion angles for the sugar-phosphate backbone and the sugar pucker were obtained using the program 3DNA (194). The overall bend-angle of the DNA was calculated using CURVES (177).

**Hardware**

HADDOCK docking runs were performed on a Transtec (Transtec AG, Tubingen, Germany) computer cluster operating with thirty-two, 2.0 Ghz, 64 bit Opteron processors. As a measure of CPU requirements, one complete run starting with 1000 structures in the rigid body docking stage could be performed in about 2 hours on 32 processors.

# 3D-DART: a DNA structure modelling server

Marc van Dijk,
Alexandre M.J.J. Bonvin

*Nucleic Acids Research*
(2009), **1**(37 web server issue):
235-239

Chapter 4

There is a growing interest in structural studies of DNA by both experimental and computational approaches. Often, 3D structural models of DNA are required, for instance, to serve as templates for homology modelling, as starting structures for macro-molecular docking or as scaffold for NMR structure calculations. The conformational adaptability of DNA when binding to a protein is often an important factor and at the same time a limitation in such studies. As a response to the demand for 3D structural models reflecting the intrinsic plasticity of DNA we present the 3D-DART server (3DNA-Driven DNA Analysis and Rebuilding Tool). The server provides an easy interface to a powerful collection of tools for the generation of DNA structural models in custom conformations. The computational engine beyond the server makes use of the 3DNA software suite together with a collection of home-written python scripts. The server is freely available at http://haddock.chem.uu.nl/dna/ without any login requirement.

## Introduction

DNA often changes its conformation as a result of interactions with various ligands; especially binding to proteins can result in large conformational changes such as helical kinks (102) or local helical untwisting (249). These play an important roll in providing complementarity to the protein binding surface and contributing to the interaction specificity (7). In order to fully understand the nature of the conformational changes taking place upon complex formation, 3D, atomic-resolution structures are required. Experimental methods such as X-ray crystallography and Nuclear Magnetic Resonance spectroscopy (NMR) but also computational approaches such as macro-molecular docking are important techniques for obtaining such 3D structures or models.

Most techniques make use of 3D structural models of DNA at some point along the structure calculation pipeline. NMR for instance can benefit from the regularity in the structure of double stranded DNA by using a model as starting point for structure calculations, thereby compensating for the lack of long range structural information. For macro-molecular docking, a starting model is often required as experimental structures might not be available. Often, starting from multiple models with different conformations improves the results. Finally, as last example, homology-modelling programs require a template model as starting point for the homology building process.

The regularity in the structure of double stranded DNA makes it especially suitable for modelling. Various software packages are available that convert a user specified base-pair sequence into a 3D structure using regular nucleotide building blocks

(85,194,201,225,316). However, most of these software packages, some of which are available via web-servers (85,225), are only able to generate models in ideal canonical conformations (194,225) or in conformations mimicking that of a free unbound structure (85).

The structures of double stranded DNA in complex with various ligands often show considerable conformational changes compared to their unbound counterparts (82,83,94,142). This plasticity originates at a "local" level in the orientation of one base relative to its Watson-Crick partner and of two base pairs relative to one another. These "local" changes accumulate and result in bending and twisting of the structure at a "global" level. Only few existing programs, such as NAMOT (316) and NAB (201), offer options to introduce custom bends in the generated DNA conformation next to giving control over all local parameters; they however require some expertise from the user and are not available as web servers. Here we describe the 3D-DART web server (3DNA-Driven DNA Analysis and Rebuilding Tool) which we developed to allow for the easy generation of 3D structural DNA models with a defined conformation by providing control over both "global" and "local" conformational features.

The generation of models is accomplished by modification of the well-established rotational and translational parameters that describe the position of one base to its Watson-Crick counterpart and of two successive base pairs relative to one another (79). It has been demonstrated in the past that rebuilding a double stranded DNA structure using these parameters results in a near native structure (194). The only exceptions are local changes in the sugar and phosphate backbone conformation.

The 3D-DART server uses the Roll, Tilt and Twist parameters to introduce bends into the structure. The other parameters can be used to "fine-tune" the conformation of the structure. Note that 3D-DART does not provide custom control of the sugar-phosphate backbone conformation (in contrast for example to NAMOT). The 3DNA software (194) is used to generate a 3D structural model from the modified parameters.

The server accepts a nucleotide sequence, a base pair (step) parameter file or a DNA-containing PDB coordinate file as input. The server returns 3D structural models with the desired conformation as well as a collection of analysis and intermediate files. Several additional and convenient functions are available to control the markup of the resulting PDB coordinate files, for instance to prepare them for use in the macro-molecular docking program HADDOCK (74,86) also developed in our group. For the same purpose the server can automatically generate a DNA restraint file (326) as an additional feature. The server is freely available at http://haddock.chem.uu.nl/dna/ without any login requirement.

## 3D-DART modelling procedure

Local bending is often at the origin of double stranded DNA distortions when in complex with various proteins (82,83,94). This type of bending can be described in terms of the vector between two successive base pairs (a base pair step). The length of this vector (Fig. 4.1, thick black arrows) describes the distance between the two base pairs in a base-pair step, also known as Rise. It usually does not vary much. In unbent canonical DNA, these vectors align with the Z-axis that represents the main helical path of the structure (Fig. 4.1A). When the DNA is bent, then the position of the vector relative to the global reference frame describes the magnitude and orientation of the bend angle. The magnitude of the bend corresponds to the vector component projected on the Y-Z plane and its orientation to the component projected on the Y-X plane (Fig. 4.1B). The accumulation of successive vectors then determines the

overall bend in the structure.

Such a bend vector can be decomposed into a Roll, Tilt and Twist base pair step parameter contributions. The 3D-DART server uses these parameters to introduce bends in the structure. The underlying algorithm is based on the transformation of the global bend vector in Euclidean space into Roll and Tilt values in the local base pair step reference frame (Fig. 4.1C). This transformation is accomplished using the following steps:

1. The definition of bend vectors requires the definition of an origin in Euclidean space set to an arbitrary base pair in the sequence (*ri*) such that the main helical path of the structure is aligned with the Z-axis. By default the central base pair is chosen (Fig. 4.1A, cyan).
2. Because the helical DNA structure is twisted the orientation component of the bend vector (*Oi*, 0°→360°) at a given base-pair step *i* needs to be corrected for the local Twist value. In figure 4.1A the red arrow illustrates the direction of the orientation component before Twist (Ω) correction and the blue one after correction. In equation 4.1 the orientation component *Oi* is corrected for the Twist value at base pair step *i* by subtracting the accumulated Twist from *ri* to *i* (Fig. 4.1A, blue circle partitions). The accumulated Twist value is positive above and negative below the reference base pair.

$$\mathrm{corO_i} = \mathrm{O_i} + \sum_{ri=1}^{i} \Omega_i \qquad \text{(Eq. 4.1)}$$

3. The bend vector with the corrected orientation component (*corOi*) at base pair step *i* is the result of the Roll (ρ) and Tilt (τ) in the local base pair step reference frame of *i* (Fig. 4.1C). The base pair tilt caused by Roll (Y'-Z' plane) is orthogonal to the base pair tilt originating from the Tilt parameter (X'-Z' plane). Together they can span 360°. Equation 4.2 defines the fractional contributions of the Roll ($fr\rho_i$) and Tilt ($fr\tau_i$) to the vectors orientation:

$$\begin{aligned} \mathrm{fr\tau_i} &= \cos(\mathrm{corO_i} \times \pi / 180) \\ \mathrm{fr\rho_i} &= \sin(\mathrm{corO_i} \times \pi / 180) \end{aligned} \qquad \text{(Eq. 4.2)}$$

The blue arrow in figure 4.1C represents the tilt fraction and the red arrow the roll fraction.

4. Finally the actual value of the Roll and Tilt parameters at base pair *i* are calculated (Eq. 4.3) to reflect both the magnitude and orientation components of the vector representing global bend angle $A_i$.

$$\begin{aligned} \rho &= \left(\sqrt{\mathrm{A_i^2} \times \left|\mathrm{fr\rho_i}\right|}\right) \times \mathrm{S_i} \\ \tau &= \left(\sqrt{\mathrm{A_i^2} \times \left|\mathrm{fr\tau_i}\right|}\right) \times \mathrm{S_i} \end{aligned} \qquad \text{(Eq. 4.3)}$$

where:
$S_i$ = 1 if $fr\rho_i$ or $fr\tau_i$ > 0
$S_i$ = -1 if $fr\rho_i$ or $fr\tau_i$ < 0
$S_i$ = 0 if $fr\rho_i$ and $fr\tau_i$ = 0

Using this algorithm the magnitude and orientation of a bend angle can be calculated for every individual base pair step in the sequence of a double stranded DNA structure. The algorithm is based on strict geometrical principles and the resulting structures are not energy minimized. Therefore it is important to emphasize that the conformation does not need to reflect an energetically favourable state. Figure 4.2 illustrates the procedure with a few examples. Introducing a 3° bend with the same orientation in a 20 base-pair structure create a smooth bend of 60° (Fig. 4.2A). Restricting a 30° bend to only the central 3 base-pair steps in the same sequence creates a kink (Fig. 4.2B). Using different angle values for every base-pair step in the sequence gives precise control over the DNA conformation (Fig. 4.2C).

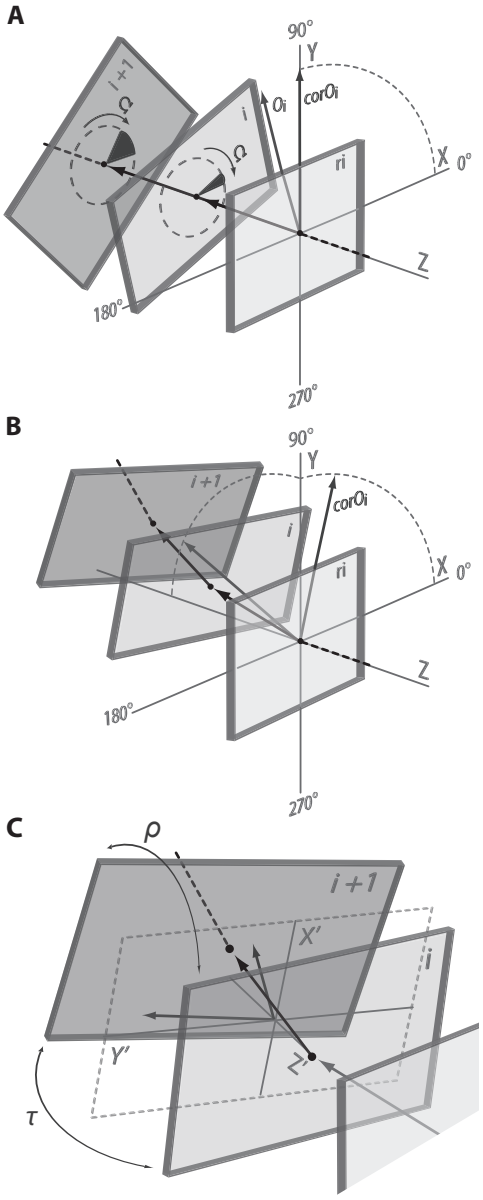



***Figure 4.1.** One block per base-pair Calladin-Drew plot of DNA illustrating the relation between the local Twist (Ω), Roll (ρ) and Tilt (τ) values and the global bend angle for a given base-pair step. Vector projections are normalized for illustrative purposes. **A**) Illustrates the correction of the orientation component of the bend vector for the local Twist value at base pair *i* and *i+1* (blue circle parts). The red arrow indicates the value of the orientation component (*Oi*) before Twist correction and the blue arrow (*corOi* aligned with Y-axis) after correction. **B**) Illustrates a bend in the structure as a result of a different bend angle vector (thick black arrows) between every successive base-pair step. The blue arrow illustrates the orientation component of the vector (Y-X plane) and the red arrow the magnitude (Y-Z plane).
**C**) Provides a detailed view or the local base-pair step reference frame between base *i* and *i+1*. The global bend vector (thick black arrow) is decomposed into a Tilt (red arrow, Y'-Z' plane) and Roll (blue arrow, X'-Z' plane) contribution.

## 3D-DART web server

### Input

The generation of custom DNA 3D structural models is based on the manipulation of the 6 base pair and 6 base pair step parameters describing the conformation of the structure. The first step in using the server is the definition of a source for the base pair (step) parameter template file that serves as a starting point for the modelling process. For this three different options are supported:

- The first option consists of inputting a nucleotide sequence defined as the bases belonging to the 5'-3' template strand. The server then uses the "fiber" module of 3DNA (194) to generate a canonical A- or B-DNA structure with the defined sequence and the "find_pair" and "analyze" module of the same software to generate the base pair (step) parameter file.
- The second input option consists of uploading an user-defined PDB coordinate file. The DNA in this file is analysed and a base-pair (step) parameter file representing its conformation is generated. This allows for the introduction of custom changes in an already existing structure.
- Finally, as third input option, the server also accepts a predefined base pair (step) parameter file in 3DNA format.

**Parameters**

The base pair (step) parameter file that results from the input data is subsequently used to start the modelling phase. At this stage, the user can introduce bends into the structure. There are two modes in which this can be accomplished referred to as "Global" and "Local":

- In the *"Global" mode* the defined bend angle is evenly distributed over all base-pair steps in the user-defined zone (Fig. 4.2A and 4.2B). The "Global" modelling mode accepts ranges of parameters so that multiple models within a given bend angle and/or bend angle orientation range can be generated. 6 models from 10° to 40° with steps of 5° for example.
- In the *"Local" mode* the bend angle and its orientation in Euclidean space can be defined uniquely for every base pair step in the user-defined zone (Fig. 4.2C).

Next to the introduction of bends the user can define custom values for the various base-pair and base-pair step parameters.


**Figure 4.2.** Two block per base Calladine-Drew plots of a 20 nucleotide B-DNA structure. The black dotted line defines the main helical path. Two examples of bending using the "Global" mode of the server are shown: **A**) illustrates a DNA conformation with a smooth bend of 60° distributed evenly over the entire structure and **B**) a conformation with a 30° kink in which the bend was restricted between base pairs 10 and 12. **C**) Shows a DNA conformation with custom angles values for every base-pair step generated using the "Local" mode of the server. The asterisk indicates the reference base pair used in the algorithm to generate the bend. The pictures were created using 3DNA (194) and PyMol (DeLano Scientific LLC, www.pymol.org).

These values are subsequently used for every base-pair or base-pair step in the sequence. If a bend is introduced then of course the Roll and Tilt values will be substituted by the ones needed to introduce the bend. If location-specific values for the parameters need to be introduced it is advised to start from a base-pair step parameter file containing these values or from a PDB coordinate file reflecting these values.

The combination of the possible input sources together with the precise modelling options makes the 3D-DART server very versatile, allowing the generation of both ideal DNA models as well as fully customized ones by introducing local conformational changes in already existing structures.

**Output**

The collection of base pair (step) parameter files resulting from the modelling phase are converted into a 3D structure in PDB format by the "rebuild" module of 3DNA. The 3D structure is built based on the nucleotide coordinates for the common bases as determined by Arnot *et al.* (50). The generated 3D structural models are returned in a zipped archive containing in addition a collection of analysis and other useful intermediate files. These include the provided input files, the base-pair (step) parameter files from which the models were generated, various bend and 3DNA analysis files and the 3D-DART log file.

In addition, the server provides a few convenience functions to further customize the output. These include options to change the PDB markup such as changing chain ID and renumbering residues. The server also offers the option to output the structures in a format and notation consistent with the macro-molecular docking program HADDOCK (74,86). In that case, an additional restraint file is generated to maintain the DNA conformation during the flexible refinement stage of the docking. These functionalities are actually used by the HADDOCK web server (http://haddock.chem.uu.nl/haddock) to automatically process DNA/RNA input structures.

# A Protein-DNA Benchmark

Marc van Dijk,
Alexandre M.J.J. Bonvin

*Nucleic Acids Research*
(2008), **36**(14 online): 1-5

Chapter 5

We present a protein-DNA docking benchmark containing 47 unbound-unbound test cases of which 13 are classified as easy, 22 as intermediate and 12 as difficult cases. The latter show considerable structural rearrangement upon complex formation. DNA-specific modifications such as flipped out bases and base modifications are included. The benchmark covers all major groups of DNA binding proteins according to the classifications of Luscombe *et al.* (197) except for the zipper-type group. The variety in test cases make this non-redundant benchmark a useful tool for comparison and development of protein-DNA docking methods. The benchmark is freely available as download from the internet.

## Introduction

Biomolecular docking has become a mature discipline within structural biology (323). Docking aims at predicting the structure of a complex given the three dimensional structures of its components. The field of protein-protein docking in particular has seen extensive progress over the last decade as witnessed by recent CAPRI (Critical Assessment of Predicted Interactions) results, a community-wide blind docking experiment (140). For protein-DNA docking, however, progress lags behind. The scarcity of information for a proper identification of interaction surfaces on DNA and its inherent flexibility have hampered the development of effective docking methods. The field of protein-DNA docking is, however, receiving increased attention and efforts are put into the development of docking methods that address the above mentioned limitations (326). Considering the importance of biomolecular interactions in system biology, gaining insight into the biochemistry of recognition and gene expression is highly relevant (260). New developments in protein-DNA docking approaches are therefore expected.

A set of well-defined test cases that form a common ground for validating and comparing the different docking methods would facilitate the development of effective protein-DNA docking methods. Such a benchmark should contain the native structures of both protein and DNA in their unbound form together with the reference structure of the complex.

We have constructed a benchmark of 47 protein-DNA test cases in a similar manner as has been done for protein-protein docking (220). The benchmark covers all major groups of protein-DNA complexes according to the classification proposed by Luscombe *et al.* (197) except for the zipper-type group. It contains a variety of challenging systems in terms of size of the interaction interface, number of individual components present in the complex and conformational changes that the unbound components undergo upon complex formation. Its diversity makes it a comparison tool for different docking methods as their performance may vary

depending on the type of complexes. This benchmark should benefit the entire docking community and offer a starting-point for the improvement of various algorithms.

## Composition of the benchmark

The protein-DNA benchmark version 1.0 (Table 5.1) contains 47 test cases. For all test cases the unbound structures of both protein and DNA are available. In addition, the reference complexes have been separated into their DNA and protein bound forms. This should allow to evaluate the performance of a docking method for bound-bound, bound-unbound and unbound-unbound cases. Although the reference structure is always from X-ray crystallography, the unbound proteins contain both solution NMR and X-ray structures. The use of an ensemble of NMR structures as starting point for the docking provides an easy way for various docking algorithms to sample additional conformational space. The benchmark contains members of all major structural groups described by Luscombe *et al.* (197) apart from the zipper-type group. These are: 16 helix-turn-helix (group 1), 3 zinc-coordinating (group 2), 5 other α-helix (group 4), 2 β-sheet (group 5), 4 β-hairpin/ribbon (group 6) and 17 enzyme (group 8) complexes.

Each test case in the benchmark poses its own challenges for a docking algorithm. A common theme throughout the benchmark are "conformational changes" either in the protein, the DNA or both. This benchmark differs from its protein-protein counterpart by the omnipresence of conformation changes. To provide some structure in the test cases we classified them as "easy", "intermediate" or "difficult". This classification is based on the interface RMSD values between the bound and unbound components of the complex:

- "easy" test case: interface RMSD between 0.0 Å and 2.0 Å
- "intermediate" test case: interface RMSD between 2.0 Å and 5.0 Å
- "difficult" test case: interface RMSD above 5.0 Å.

*An "easy" test case*

The individual components from this group of complexes do not change significantly the conformation of their interface upon binding. Conformational changes at the interface of the protein are mostly brought about by small flexible loop rearrangements. This does not mean that the components can always be regarded as rigid. Conformational changes at the interface of the DNA often cause the DNA to bend and twist in the interface region (see DNA RMSD values in Table 5.1). A representative example from this group is the Papillomavirus replication initiation domain E-1 (PDB entry 1ksy, Fig. 5.1A).

*An "intermediate" test case*

Unbound components of this group undergo more pronounced structural rearrangements in their interface upon complex formation. The type of conformational changes involves global and local domain rearrangements in the protein and global conformational change in the DNA. An example is the intron-encoded homing endonuclease I-PpoI complex (PDB entry 1a73, Fig 5.1B), the protein shows little conformational change upon binding but the DNA is heavily kinked in its center.

*A "difficult" test case*

In the difficult cases the extent of structural rearrangement upon complex formation increases even further. In addition to the conformational changes occurring in the "intermediate" test cases, the "difficult" group contains complexes with features like structural transitions and major domain reorientations in the protein. An example is the proline utilization transcription activator (PDB entry 1zme, Fig. 5.1C), a protein that has two DNA interaction domains linked

together by a long highly flexible loop; the dimerisation interface connecting the two DNA interaction domains show a loop to sheet transition upon DNA binding. In the PVUII endonuclease complex (PDB entry 1eyu, Fig. 5.1D) the individual protein chains do not show much conformational changes but a hinge point connecting them facilitates a "clamping" motion upon binding. This results in a large RMS deviation between bound and unbound structures. This is an example of global domain motions upon binding.

The benchmark also contains several structures with special features such as strand breaks (PDB entries 1g9z, 103t and 3bam) and flipped out bases in the DNA (PDB entries 1diz, 1emh, 1vas and 7mht).

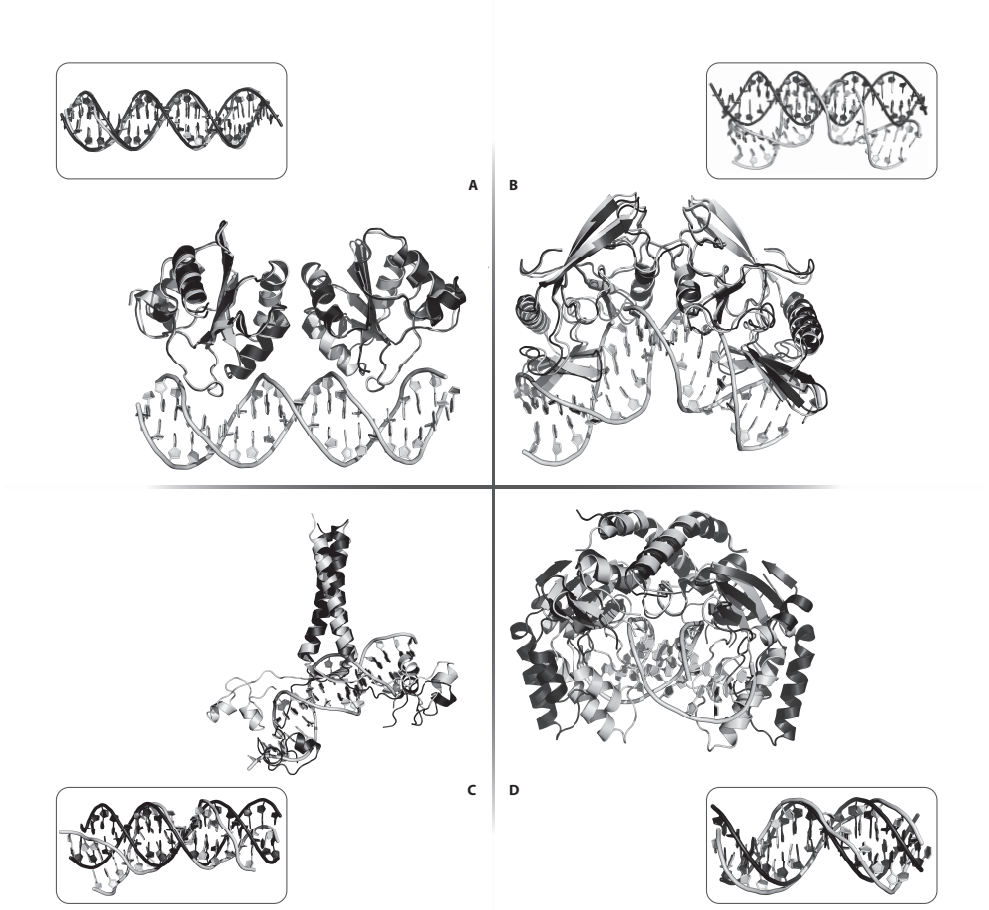



***Figure 5.1.** Illustration of "easy" (interface RMSD < 2.0Å), "intermediate" (2.0Å ≤ interface RMSD < 5.0Å) and "difficult" (interface RMSD ≥ 5.0Å) test cases from the protein-DNA benchmark. "Easy" test case: the Papillomavirus replication initiation domain E-1 (PDB id 1ksy) (interface RMSD = 1.6Å) (**A**). "Intermediate" test case: the intron-encoded homing endonuclease I-PpoI complex (PDB id 1a73) (interface RMSD = 4.3Å) (**B**). "Difficult" test cases: the proline utilization transcription activator (PDB id 1zme) (interface RMSD = 5.8Å) (**C**) and the PVUII endonuclease complex (PDB id 1eyu) (interface RMSD = 6.8Å) (**D**). The bound form of the complex is shown in yellow and the unbound protein in blue. The bound- and canonical B-form DNA structures are shown as insets to highlight the conformational changes in the DNA.

We constructed this benchmark as a test base to stimulate developments in the field of protein-DNA docking and will use it in particular for further developing our own protein-DNA docking approach (326). Ideally, the classification of "easy", "intermediate" or "difficult" could have been based on docking results; at this stage, however, we chose to purely base it on conformational changes as measured by the RMS deviations between bound and unbound form. Basing the classification on HADDOCK results would have introduced a bias not only toward the amount of conformational changes, but also toward our ability to predict protein-DNA interfaces since HADDOCK requires some kind of input to drive the docking process. We will of course proceed with evaluating our performance on this benchmark, but this is outside the scope of this chapter.

In conclusion, allowing for structural rearrangements in both protein and DNA during docking while maintaining the helical character of DNA is a major challenge in protein-DNA docking. The large variety of protein-DNA complexes in the benchmark should provide a valuable test set to evaluate and improve docking algorithms. Version 1.0 of the benchmark is available from the web site: http://haddock.chem.uu.nl/dna/benchmark.html

## Materials and Methods

### RCSB Protein Data Bank query

A non-redundant benchmark was generated from structures deposited in the RCSB Protein Data Bank (PDB) (28). The PDB (as of September 2007) was queried for all entries containing X-ray crystallographic structures with a resolution better than 3.0 Å containing both protein and DNA. Complexes containing DNA structures with a sequence length smaller than 8 base pairs and protein structures containing mutations in the core and or interface region were removed.

For the resulting complexes the PDB was queried for unbound protein entries. Structures resolved using NMR or X-ray crystallography with a resolution better than 3.0 Å were retrieved. Structures with a sequence similarity larger than or equal to 90% were removed. Structures were regarded as redundant if the raw alignment score is positive, more than 80% of their sequences are aligned and more than 60% of the sequences are identical. Sequence alignments were performed using the Needleman-Wunsch algorithm as implemented in the LSQMAN software package (293) with a gap penalty of 5.

### Generation of unbound DNA models

Models for unbound DNA were generated using the DNA analysis and rebuilding program 3DNA (194) with the base-pair sequence of the DNA in the reference complex. The models were generated in canonical B-DNA conformation (fiber model 4) using the nucleotide building blocks as determined in the fiber diffraction studies of Arnott *et al.* (50). Structures with overhanging base-pairs were converted to all-paired structures by adding their Watson-Crick counterparts.

### Structure post-processing

The residue numbering of the bound and unbound components was matched to allow for easy comparison. The DNA was assigned one chain identifier and renumbered. Structures of unbound proteins that contain more than one chain were assigned a single chain identifier instead of being separated into their individual components; residues were renumbered to avoid overlap in numbering. Atom and residue names were matched to the topallhdg5.3.pro (188) and dna-rna_allatom.top topology files (41) naming for direct use in HADDOCK (86).

**Analysis**

The size of the interaction interface between protein and DNA is expressed in terms of the Buried Surface Area (BSA, Table 5.1) of the DNA in the complex. The BSA was calculated using NACCESS (Hubbard, S.J., Thornton, J.M. 1993) with a probe radius of 1.4 Å. The conformational changes between the unbound and the bound states are expressed in terms of the Root Mean Square Deviation (RMSD) calculated using ProFit (Martin, A.C.R., http://www.bioinf.org.uk/software/profit/). These were calculated in three different ways:

1. Conformational change of the protein-DNA interface was calculated by superimposition of all Cα and phosphate atoms at the interface. Residues belonging to the interface are identified as those having atoms within 5.0 Å intermolecular distance of one another (RMSD Inter., Table 5.1). The interface RMSD values were used to classify the test cases as "easy", "intermediate" or "difficult" (see below).
2. As the conformational change in the DNA tends to affect the complete molecule the RMSD of the DNA was calculated by superimposition of all phosphate atoms (RMSD DNA, Table 5.1).
3. Conformational changes in the protein such as global domain reorientations and flexible segments not located at the interface are represented by means of the RMSD calculated over all Cα atoms of the protein (RMSD Prot, Table 5.1).

**Table 5.1.** The protein-DNA benchmark

| Complex | | Protein | | DNA | | | RMSD | | |
|---|---|---|---|---|---|---|---|---|---|
| PDB id[a] | Cat.[a] | PDB id[a] | Description | Sequence 5'-3' [c] | Nr.[d] | BSA[e] | Inter.[f] | DNA[g] | Prot[h] |
| **"Easy" targets** | | | | | | | | | |
| 2c5r | 1 | 2bnk[X] | Phage PHI29 replication organizer protein P16.7 | TCCACCGG | 4 | 402 | 0.49 | 0.49 | 0.82 |
| 1pt3 (A:C:D) | 8 | 1mo8[X] | Col-E7 nuclease domain | GCGATCGC | 2 | 730 | 1.35 | 2.09 | 1.36 |
| 1mnn | 1 | 1mn4[X] | Sporulation specific transcription factor NDT80 | TGCGACACAAAAACT | 2 | 1292 | 1.48 | 1.81 | 0.83 |
| 1fok | 1 | 2fok[X] | Restriction endonuclease FOKI | TCGGATGATAACGCTAGTCAT | 2 | 1920 | 1.53 | 2.51 | 1.09 |
| 1ksy (A:C:D:F) | 4 | 1fo8[X] | Papillomavirus replication initiation domain E-1 | ATAATTGTTGTCAACAATTAT | 3 | 1020 | 1.58 | 2.56 | 0.52 |
| 3cro | 1 | 1zug[N] | Phage 434 CRO | AAGTACAAACTTTCTTGTAT | 3 | 1473 | 1.58 | 2.66 | 1.17 |
| 1emh | 8 | 1akz[X] | Human uracil-DNA glucosylase | TGT(P2U)ATCTTT | 2 | 869 | 1.62 | 4.53 | 1.46 |
| 1h9t | 1 | 1e2x[X] | FADR, fatty acid responsive transcription factor | CATCTGGTACGACCAGATC | 3 | 1622 | 1.68 | 3.88 | 0.77 |
| 1tro (A:C:I:J) | 1 | 3wrp[X] | TRP repressor | TGTACTAGTTAACTAGTACA | 3 | 1540 | 1.70 | 3.08 | 1.42 |
| 1by4 (A:B:E:F) | 2 | 1rxr[N] | Retinoid X receptor DNA binding domain | TAGGTCAAAGGTCAG | 3 | 1480 | 1.77 | 1.46 | 2.23 |
| 1hjc (A:B:C) | 5 | 1ean[X] | RUNX1 runt domain | GAACTCTGTGGTTGCGG | 2 | 634 | 1.80 | 2.88 | 0.97 |
| 1diz (A:E:F) | 8 | 1mpg[X] | e.coli 3-methyladenine DNA glycosylase II | TGACATGA(NRI)TGCCT | 2 | 805 | 1.82 | 5.80 | 0.46 |
| 1rpe | 1 | 1r63[N] | Phage 434 repressor | ACAAACAAGATACATTGTATA | 3 | 1430 | 1.87 | 2.97 | 0.94 |
| **"Intermediate" targets** | | | | | | | | | |
| 1vrr | 8 | 1sdo[X] | Restriction endonuclease BSTYI | TTATAGATCTATAA | 3 | 2098 | 2.08 | 2.11 | 2.22 |
| 1f4k | 1 | 1bm9[X] | Replication terminator protein | CTATGAACATAATGTTCATAG | 3 | 1741 | 2.26 | 1.94 | 2.29 |
| 1k79 (A:B:C) | 1 | 1gvj[X] | ETS-1 DNA binding and autoinhibitory domain | TAGTGCCGGAAATGTG | 2 | 912 | 2.37 | 3.82 | 0.80 |
| 1kc6 (A:B:E:F) | 8 | 2aud[X] | Restriction endonuclease HINCII | CCGGTCGACCGG | 3 | 2658 | 2.38 | 4.67 | 1.38 |

**Table 5.1.** Continued

| Complex | | Protein | | DNA | | | RMSD | | |
|---|---|---|---|---|---|---|---|---|---|
| PDB id[a] | Cat.[a] | PDB id[a] | Description | Sequence 5'-3' [c] | Nr.[d] | BSA[e] | Inter.[f] | DNA[g] | Prot[h] |
| 1ea4 (D:E:F:G:W:X) | 6 | 2cpg[X] | Transcription repressor COPG | TAACCGTGCACTCAATGCAATC | 3 | 1473 | 2.43 | 4.48 | 0.64 |
| 1z63 (A:C:D) | 8 | 1z6a[X] | Sulfolobus solfataricus SWI2/SNF2 ATPase core domain | ATTGCCGAAGACGAAAAAAA | 2 | 603 | 2.51 | 2.74 | 2.27 |
| 1r4o | 2 | 1gdc[N] | Glucocorticoid receptor | CCAGAACATCGATGTTCTGT | 3 | 1401 | 2.61 | 3.05 | 1.91 |
| 1azp | 6 | 1sap[N] | Hyperthermophile chromosomal protein SAC7D | GCGATCGC | 2 | 778 | 2.70 | 3.77 | 2.76 |
| 1w0t | 1 | 1ba5[N] | HTRF1 DNA-binding domain | CTGTTAGGGTTAGGGTTAGA | 3 | 1545 | 2.78 | 3.20 | 2.47 |
| 1cma | 6 | 1mjk[X] | Methionine repressor | TTAGACGTCT | 2 | 775 | 2.81 | 2.60 | 2.05 |
| 1jj4 | 4 | 1f9f[X] | Papillomavirus type 18 E2 | CAACCGAATTCGGTTG | 2 | 1169 | 2.83 | 3.32 | 2.25 |
| 1vas | 8 | 1eni[X] | T4 pyrimidine dimer specific excision repair | ATCGCGTTGCGCT | 2 | 1445 | 3.04 | 6.99 | 1.42 |
| 4ktq | 8 | 1ktq[X] | DNA polymerase I | GACCACGGCGC(DOC) | 2 | 1685 | 3.23 | 3.64 | 1.97 |
| 1z9c (A:C:D) | 1 | 1z91[X] | Organic hydroperoxide resistence transcription regulator | TACAATTTAATTGTATACAATTTAATTGTA | 3 | 2107 | 3.24 | 4.26 | 4.18 |
| 1ddn | 1 | 2tdx[X] | Diphtheria TOX repressor | ATATAATTAGGATAGCTTTACCTAATTATT TTAA | 5 | 2877 | 3.26 | 7.25 | 0.50 |
| 2irf | 1 | 1irg[N] | Interferon Regulatory Factor 2 | AAGTGAAAGUGA | 2 | 898 | 3.35 | 2.23 | 3.83 |
| 1jto | 1 | 1jus[X] | Multidrug binding transcription factor QACR | CTTATAGACCGATCGATCGGTCTATAAG | 2 | 2484 | 3.49 | 4.58 | 3.53 |
| 1g9z | 8 | 207m[X] | I-CreI endonuclease | GCAAAACGTCGTGAGACAGTTTCG | 2 | 3255 | 3.67 | 5.02 | 4.21 |
| 1a73 | 8 | 1evx[X] | Intron-encoded homing endonuclease I-PpoI | TTGACTCTCTTAAGAGAGTCA | 2 | 2076 | 4.26 | 8.22 | 1.20 |
| 2fio | 4 | 2fib[X] | Phage PHI29 transcription regulator P4 | AAAAACGTCAACATTTTATAAAAAAGTC TTGCAAAAAGT | 2 | 1114 | 4.41 | 8.03 | 0.67 |
| 1qne (A:C:D) | 5 | 1vok[X] | Adenovirus major late promotor TBP | GCTATAAAAGGGCA | 2 | 1487 | 4.57 | 8.54 | 0.89 |
| 1zs4 | 1 | 1zpq[X] | Phage lambda CII | CCTCGTTGCGTTTGTTTGCACGAAT | 2 | 1358 | 4.71 | 2.97 | 3.77 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **"Difficult" targets** | | | | | | | | | |
| 1qrv | 4 | 1hma$^{N}$ | High mobility group protein D | GCGATATCGC | 3 | 1204 | 5.19 | 7.68 | 3.91 |
| 103t | 1 | 1g6n$^{X}$ | CAP-CAMP | GCTTTTTACGCTAGATCTAGCGTAAAAA GCGC | 2 | 1277 | 5.20 | 10.6 | 2.55 |
| 1b3t | 4 | 1vhi$^{X}$ | Epstein-Barr virus nuclear antigen-1 | GGAAGCATATGCTTCCC | 2 | 2627 | 5.32 | 3.91 | 3.53 |
| 3bam | 8 | 1bam$^{X}$ | Restriction endonuclease BAMHI | TATGGATCCATA | 3 | 2208 | 5.55 | 2.19 | 4.50 |
| 1rva | 8 | 1rve$^{X}$ | Eco RV endonuclease | AAAGATATCTTT | 2 | 2350 | 5.68 | 9.78 | 3.88 |
| 1zme | 2 | 1ajy$^{N}$ | Proline utilization transcription activator PUT3 | ACGGGAAGCCAACTCCGT | 2 | 1362 | 5.76 | 4.68 | 8.64 |
| 1dfm | 8 | 1es8$^{X}$ | Restriction endonuclease BGLII | TATTATAGATCTATAAAT | 3 | 2735 | 6.31 | 3.04 | 4.68 |
| 1bdt | 6 | 1arq$^{N}$ | Phage P22 Arc gene regulating protein | TATAGTAGAGTGCTTCTATCATT | 3 | 2109 | 6.45 | 4.90 | 5.20 |
| 7mht | 8 | 2hmy$^{X}$ | HHAI methyltransferase | GTCAGCGCATGG | 2 | 1613 | 6.71 | 2.55 | 3.84 |
| 2fl3 | 8 | 1ynm$^{X}$ | Restriction endonuclease HINP1I | CCAGCGCTGG | 2 | 1670 | 6.71 | 2.95 | 4.37 |
| 1eyu | 8 | 1pvu$^{X}$ | PVUII endonuclease | TGACCAGCTGGTCA | 2 | 2068 | 6.82 | 4.49 | 6.36 |
| 20aa | 8 | 20a9$^{X}$ | Restriction endonuclease MVAI | GGTACCTGGATG | 2 | 2009 | 8.95 | 8.15 | 8.02 |

a) The RCSB PDB accession number for the structures used. Specific chains are in parenthesis. Structures for the unbound protein were either solved by X-ray crystalography (x) or NMR spectroscopy (N).

b) The classification of the protein-DNA complexes in 8 different groups according to the scheme of Luscombe *et al* (1).

c) The base sequence of the DNA in the bound complex also used for generating the unbound DNA structure. Some sequences contain modified bases. These are: DOC (2',3'-dideoxycytidine-5'-monophosphate), NRI (phosphoric acid mono-(4-hydroxy-pyrrolidin-3-ylmethyl) ester) and P2U (2'-deoxy-pseudouridine-5'monophosphate)

d) The number of individual biomolecules that need to be docked to reconstruct the complex.

e) Buried surface area of the DNA upon complex formation in Å$^{2}$.

f) The RMSD (Å) from the bound form calculated over the interface C$\alpha$ and phosphate atoms of the unbound protein structure after superposition onto the reference complex.

g) The RMSD (Å) from the bound form calculated over all phosphate atoms of the unbound DNA after superposition onto the reference complex.

h) The RMSD (Å) from the bound form calculated over C$\alpha$ atoms of the unbound protein after superposition onto the reference complex.

# Pushing the limits of what is achievable in protein-DNA docking.

# Benchmarking the performance of HADDOCK

Marc van Dijk
Alexandre M.J.J. Bonvin

*Manuscript submitted for publication*

The intrinsic flexibility of DNA and the difficulty in identifying its interaction surface have long been challenges that prevented the development of efficient protein-DNA docking methods. We have demonstrated before [M. van Dijk et al. (2006) *Nucleic Acids Res*. 34 (11), 3317-3325] the ability of the data driven docking method HADDOCK to deal with these challenges by an explicit flexibility treatment and the use of custom-built DNA structural models. In this study we improve our method and put it to the test on a set of 47 complexes from the protein-DNA docking benchmark. We show that HADDOCK is able to predict many of the specific DNA conformational changes required to assemble the interface(s) using both ideal and experimentally derived restraints to drive the docking. Our DNA analysis and modelling procedure is able to capture the bend and twist motions occurring upon complex formation and use these to generate an ensemble of custom-built DNA structural models, more closely reassembling the bound form, for use in a second docking round. Top ranking solutions throughout the benchmark readily score one and two stars according to CAPRI quality criteria, achieving an overall success rate of 94% acceptable solutions. Our improved protocol makes it possible to successfully predict even the challenging protein-DNA complexes in the benchmark from their unbound components. Finally, our method is the first to readily dock multiple molecules (N > 2) simultaneously, pushing the limits of what is currently achievable in the field of protein-DNA docking.



## Introduction

The computational docking field is proceeding ever faster to becoming an integral part of the research workflow in life sciences. Most of the developments in docking methodology were pioneered in the fields of small molecule docking and protein-protein docking (120,261,281). Docking has become a valuable tool in drug design, molecular interaction studies, NMR and X-ray structural studies, biochemical experiment design and validation (111,146,167). While docking is flourishing in these fields, less progress has been made in the development of successful protein-DNA docking algorithms. This is in part due to two system-dependent problems: 1) identifying the location of the interaction interface(s) on the DNA and 2) modelling DNA conformational changes while maintaining a correct representation of the DNA double helix during a simulation. The

field of protein-DNA docking is, however, receiving renewed interest as the vital role of protein-DNA interactions in regulating gene expression and guarding genome integrity has become imminent (90). As a consequence new protein-DNA docking methods are put forward and proven protein-protein docking concepts are extended to deal with these systems (1,8,20,98,99,156,191,252,262,270).

We have in the past adapted our data driven docking method HADDOCK, to deal with protein-DNA systems (326) and showed that it is able deal with the two main challenges mentioned above. The ability of HADDOCK to use experimental data to drive the docking greatly facilitates the identification and positioning of the interaction interfaces during the docking (216,323). The incorporation of flexibility, both explicitly during the docking and implicitly by the use of custom-built DNA structural models, has proven to facilitate the conformational changes in the protein and DNA needed to establish the complex. The protocol was initially tested by docking the unbound structures of three monomeric transcription factors to their respective operator half-sides (phage 434 Cro (222), phage λ Arc (276) and Escherichia Coli Lac (60)). The resulting near native docking solutions reproduced many of the contacts observed in the experimental structures as well as specific conformational changes in the DNA. Our initial protein-DNA docking protocol has been successfully used in a number of practical applications by various laboratories worldwide (31,44,110,189,294). Driven by this success we have worked on improving the method's performance and user friendliness by facilitating the generation of custom DNA structural models (328) as well as establishing a protein-DNA docking benchmark as a test bed for future developments (327). Next to that, HADDOCK has been made available to the community as a web server (http://www.haddocking.org; http://haddock.chem.uu.nl).

Here we bring all these elements together and challenge our method using the 47 test cases from the protein-DNA benchmark to define the limits of our current approach. We focus on the same two questions addressed in chapter 3: How successful is the method in dealing with conformational changes upon complex formation and how well is it able to identify the correct interaction interfaces? Compared to the three test cases used previously, the 47 test cases in the benchmark pose some considerable challenges. The initial test cases were all major groove interacting transcription factors in their monomeric form, targeting one operator half-side that effectively spans one helical turn of DNA. The DNA-interacting domain of these transcription factors only changes conformation with respect to the side-chains of the DNA-interacting residues. The global conformational changes in the DNA were expressed as a uniform bend and change in groove width. In contrast, among the 47 test cases of the benchmark, not only transcription factors but also enzymes and structural proteins are present. These interact using a variation of structural domains, often involving multiple proteins, targeted to one or multiple sites on the DNA. Furthermore, the DNA length is often more than one helical turn. As a consequence, conformational changes can no longer be expressed in a smooth and uniform way but rather as an accumulation of local DNA bending and twisting events. To cope with these challenges we have improved our method for the generation of custom DNA structural models by extending its ability to capture the main bend and twist motions occurring in the DNA upon complex formation, and by subsequently using this information for the generation of custom DNA models.

The results, again, show that the use of explicit flexibility in combination with

implicit flexibility by means of an ensemble of custom-built DNA structural models, greatly improves the protein-DNA docking efficiency with respect to rigid-body docking. This is especially clear for the intermediate and difficult categories of the benchmark where DNA conformational changes readily occur. The use of experimental information for the docking of a representative subset of the benchmark, demonstrates the ability of our method to identify the correct interfaces and assemble the complex under "real life" docking conditions. Furthermore, our method is the first to dock multiple molecules simulations, a valuable feature in a benchmark containing 40% of multi-component complexes. Top ranking docking solutions throughout the benchmark readily score one and two stars according to the CAPRI quality criteria (139) and three star predictions are getting within reach for "easy" test cases.

To our knowledge this is the first time a protein-DNA docking study of such a magnitude has been performed. Our results stress the importance of conformational adaptation in the docking of protein-DNA complexes and show the potential of HADDOCK to deal with them. We hope that they will stimulate the docking community to put their methods to the test on the same benchmark and foster further developments.



## Results

The power of HADDOCK as a method relies among others on its use of Ambiguous Interaction Restraints (AIRs) and explicit flexibility. An AIR defines that a residue on the surface of a biomolecule should be in close vicinity to another residue or group of residues on the partner biomolecule when they form the complex. By default this is described as an ambiguous distance restraint between all atoms of the source residue to all atoms of all target residue(s) that are assumed to be in the interface in the complex. The effective distance between all those atoms, $d_{iAB}^{eff}$ is calculated as follows:

$$d_{iAB}^{eff} = \left( \sum_{m_{iA}=1}^{N_{Aatom}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{Batom}} \frac{1}{d_{m_{iA} n_{kB}}^{6}} \right)^{-1/6} \quad \text{(Eq. 6.1)}$$

Here $N_{Aatom}$ indicates all atoms of the source residue on molecule A, $N_{resB}$ the residues defined to be at the interface of the target molecule B, and $N_{Batom}$ all atoms of a residue on molecule B. The $1/r^6$ summation somewhat mimics the attractive part of the Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the biomolecules are in contact. The AIRs are incorporated as an additional energy term to the energy function that is minimized during the docking. The ambiguous nature of these restraints easily allows experimental data that often provide evidence for a residue making contacts to be used as driving force for the docking. As such the AIRs define a network of restraints between the possible interaction interface(s) of the molecules to be docked without defining the relative orientation of the molecules, minimizing the necessary search through conformational space needed to assemble the interfaces. Because the AIRs are part of the energy function they might also contribute to inducing the conformational changes during the flexible stage of the docking.

To objectively answer the question: "how successful is HADDOCK in dealing with conformational changes upon complex formation?" the effects of the quality and quantity of AIRs on complex formation and conformational change should be kept to a minimum. This was realized by constructing ideal AIR restraint sets based on the true interface(s) of the reference complexes (see Materials and Methods). Using these restraints we first evaluated the ability of HADDOCK to reconstruct the complex from its components in their bound conformation. Challenges in reconstruction due to

structural characteristics, the inability of the restraints to drive correct complex formation or selection of top ranking solutions due to scoring problems can be identified at this stage. Next we used the same restraints to drive the docking between the unbound protein and a canonical B-DNA 3D structural model using our two-stage protein-DNA docking approach. We focused on the two stages individually, first evaluating the effects of explicit flexibility on the docking by comparing the docking solutions from rigid body refinement with those after semi-flexible refinement. Subsequently we analyzed the conformation of the DNA in the final docking solutions. Here, the focus was on the ability of HADDOCK to introduce those specific DNA conformational changes in terms of DNA bending and twisting that can lead to the final conformation of the DNA in the complex. With this information an ensemble of custom DNA structural models was generated using a modified protocol of our 3D-DART DNA modelling web server (see Materials and Methods). The resulting models were used as input for a second, "refinement", docking run. The results were compared with those of the previous run starting from a canonical B-DNA structural model to analyze the effect of this implicit treatment of flexibility. Finally the same two-stage docking protocol was applied to a subset of six test cases from the benchmark using AIR restraints based on experimental information obtained from literature sources.

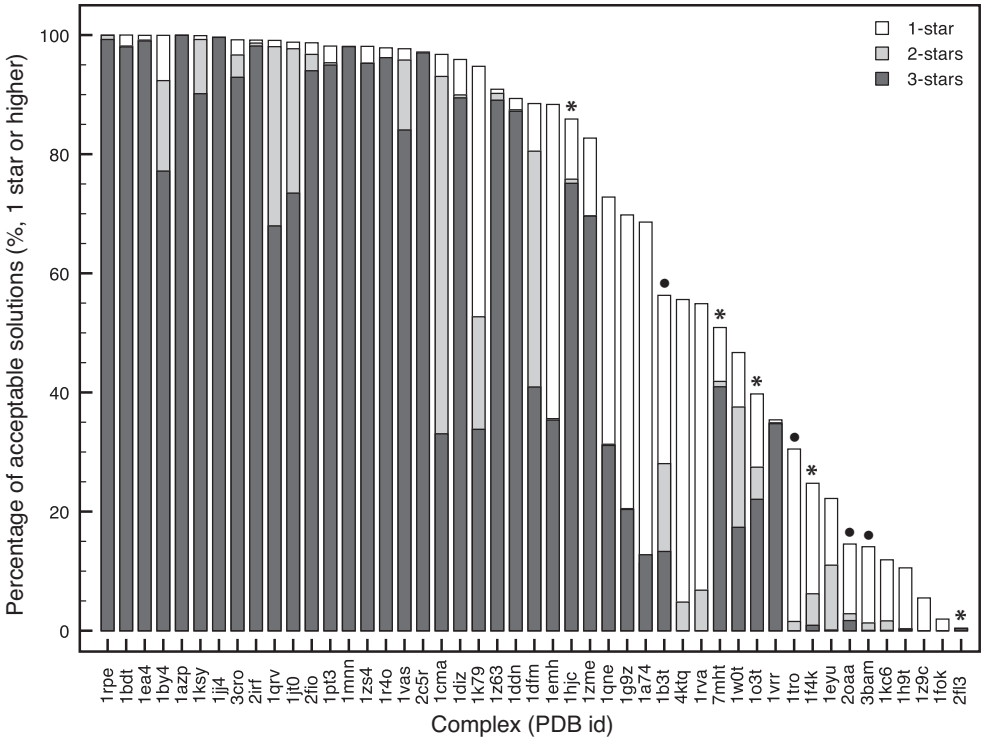



**Figure 6.1.** Cumulative bar graph expressing the quality of the docking solutions according to the CAPRI star rating for 2000 bound-bound rigid-body docking solutions. Complexes are sorted according to the total number of obtained stars. The non-bonded interactions were scaled down for the complexes marked with an asterisk (*, scaling factor 0.01) or filled circle (•, scaling factor 0.00001).

**Bound-bound docking**

A bound-bound docking experiment is essentially an exercise of separating the reference complex into its individual biomolecules and reconstructing it again. As the different components are already in their bound conformation flexibility is not required and only rigid-body docking needs to be performed. The ability of HADDOCK to sample conformational space in search of the correct interaction interface(s) using ideal AIR restraints was evaluated using the CAPRI star ranking as a quality measure commonly used in protein-protein docking (139). These criteria define one-star predictions as 'acceptable', two-star as 'medium' and three-star as 'high' quality with respect to their reference structure (see Materials and Methods).

The results illustrate that for the majority of the test cases high quality, three star, solutions are generated (Fig. 6.1). For the first half of the test cases (left half of Fig. 6.1) more than 95% of the solutions ranked one star or higher but for the remaining a sharp decline in the total number of acceptable solutions was observed. The latter group of test cases corresponds mostly with the "intermediate" and "difficult" categories of the benchmark. They have larger and more segmented interface(s). Many of them require close interpenetration of the protein-DNA and/or protein-protein interfaces to generate a well-packed complex. These involve enzymes that perform their catalytic function on single nucleotides that are flipped out of the helix into a catalytic pocket of the protein (1emh, 7mht) or, in the remainder of the cases, multiple DNA binding proteins for which the protein-DNA and/or protein-protein interfaces are composed of a number of interpenetrating loops and secondary structure elements.

Effective docking of a number of these cases is hindered by non-bonded repulsions associated with interface penetration and the correct alignment of the segmented interfaces during the rotation and translation stages of the rigid body refinement. Several of the test cases described above yielded very few to no acceptable solutions using the default docking protocol. A significant improvement in docking efficiency was realized by scaling down the non-bonded energy terms during the docking allowing interpenetration to occur (Fig. 6.1, asterisk- and filled circle -marked complexes). Despite the differences in total number of acceptable solutions, the 10 best solutions selected based on the HADDOCK score in all cases coincided with the best solutions based on the CAPRI criteria. This indicates that the HADDOCK scoring function at this stage is sufficient to retrieve the best solutions.

**Unbound-unbound docking starting from canonical B-DNA models**

We proceeded with the docking of the unbound conformation of the proteins with canonical B-DNA models using ideal AIRs. To increase the sampling of conformational space for the proteins, especially those that use flexible loops to interact with DNA grooves, we first performed a simulated annealing on the interface residues followed by a refinement in explicit water. This procedure resulted in an ensemble of 5 structures, including the original unbound protein, sampling different conformations of the interface. The protein-DNA docking protocol, at this stage, effectively incorporates two modes of flexibility: implicit sampling by means of the ensemble of protein starting structures and explicit sampling of protein and DNA conformational space during semi-flexible refinement.

Figure 6.2 illustrates the docking results using only rigid-body docking and the effect of a subsequent semi-flexible refinement. Here, the grey plus the white bars indicate the percentage of one and/or two star solution obtained after rigid-body docking. The grey bars alone indicate the percentage after semi-flexible refinement therefore

corresponding to a reduction in the number of acceptable solutions. In case of a black bar, the white bars alone indicate the percentage of one and/or two stars after rigid-body docking and the white plus the black bars indicate the percentage after semi-flexible refinement therefore corresponding to an improvement. For a number of complexes, acceptable and medium quality solutions (one and two stars) were already obtained after rigid-body docking. In all cases, except for 1dfm, the number of acceptable or higher solutions increased significantly after semi-flexible refinement (black bars). The number of acceptable and higher solutions clearly divides the complexes into three groups that coincide reasonably well with the "easy", "intermediate" and "difficult" categories of the benchmark. For the "easy" category the inclusion of explicit flexibility readily results in a shift from one star to two star solutions, for the "intermediate" category the number of one star solutions greatly improves and for the "difficult" category one star solutions are

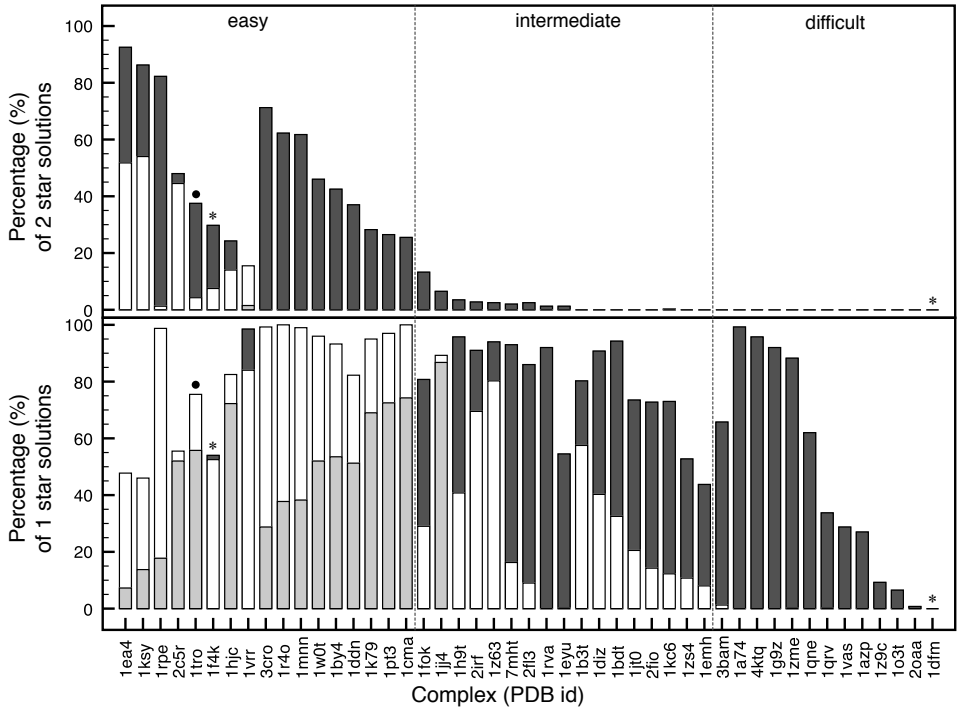



**Figure 6.2.** Effect of explicit flexibility on the docking results expressed in terms of CAPRI one-star (bottom panel) and two-star (top panel) solutions for the unbound docking starting from a canonical B-DNA model. The grey plus the white bars indicate the percentage of one and/or two star solution obtained after rigid-body docking. The grey bars alone indicate the percentage after semi-flexible refinement, therefore corresponding to a reduction. In case of a black bar, the white bars alone indicate the percentage of one and/or two stars after rigid-body docking and the white plus the black bars indicate the percentage after semi-flexible refinement, therefore corresponding to an improvement compared to rigid-body docking. The non-bonded interactions were scaled down for the complexes marked with an asterisk (*, scaling factor 0.01) or filled circle (•, scaling factor 0.00001). Complexes are sorted according to the number of one and two-star solutions obtained after semi-flexible refinement. This sorting results in a reclassification of the benchmark into "easy", "intermediate" and "difficult" categories.

only achieved because of explicit flexibility. As in the case of bound-bound docking, a few complexes could only be effectively docked by scaling down the non-bonded interactions (Fig. 6.2, asterisk and filled circle marked complexes).

**Unbound-unbound docking starting from custom-build B-DNA structural models**

The previous docking results show the improvements that can be obtained when using explicit flexibility versus rigid-body docking. In all cases, the DNA and the proteins could adapt their conformation to better interact with each other. For the DNA, these conformational changes range from small local changes in helical bend and groove width, while maintaining a relative straight helix, to larger global changes that effectively bend en twist the DNA structure. However, the amount of conformational space that can be sampled during the semi-flexible refinement stage is limited. Starting from a canonical B-DNA structural model, the semi-flexible refinement stage improved the DNA model on average by 0.84 ± 0.36 Å all heavy atom RMSD with respect to the target. This clearly cannot account for the

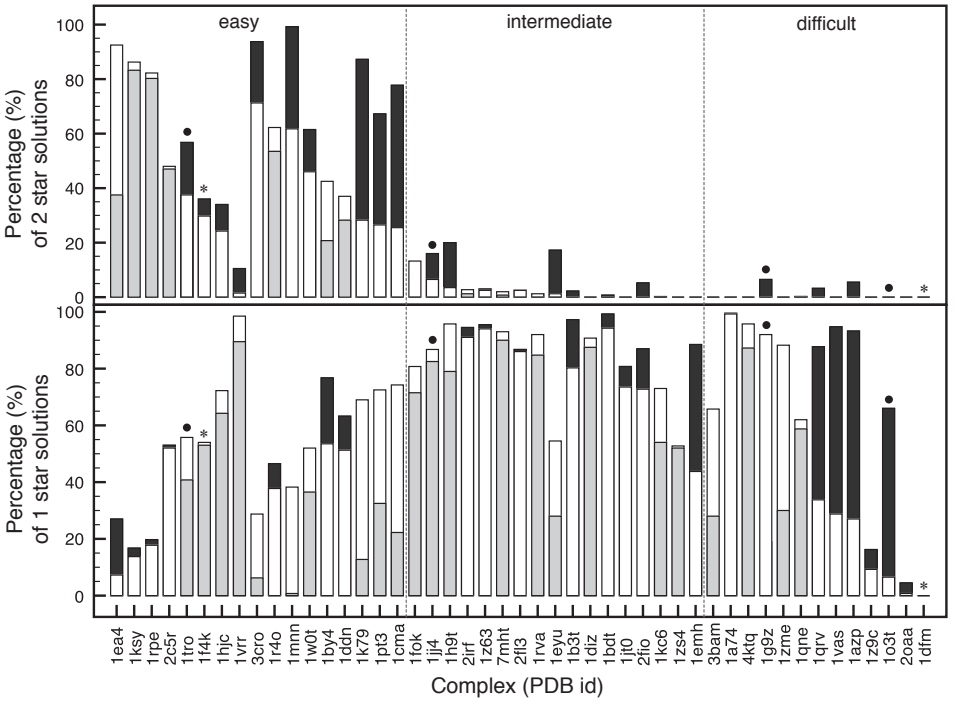



**Figure 6.3.** Effect of an ensemble of custom DNA 3D structural models on the docking results expressed in terms of CAPRI one-star (bottom panel) and two-star (top panel) solutions. The grey plus the white bars indicate the percentage of one and/or two star solutions obtained after the first docking run using canonical B-DNA as starting structure (Figure 6.2, final result). The grey bars alone indicate the percentage when using the custom-built DNA structure ensemble, therefore corresponding to a reduction. In case of a black bar, the white bars alone indicate the percentage of one and/or two stars after the first docking run and the white plus the black bars indicate the percentage when using the custom-built DNA structure ensemble, therefore corresponding to an improvement due to the use of custom-built DNA models. The non-bonded interactions were scaled down for the complexes marked with an asterisk (*, scaling factor 0.01) or filled circle (•, scaling factor 0.00001). Complexes are sorted as in Figure 6.2.

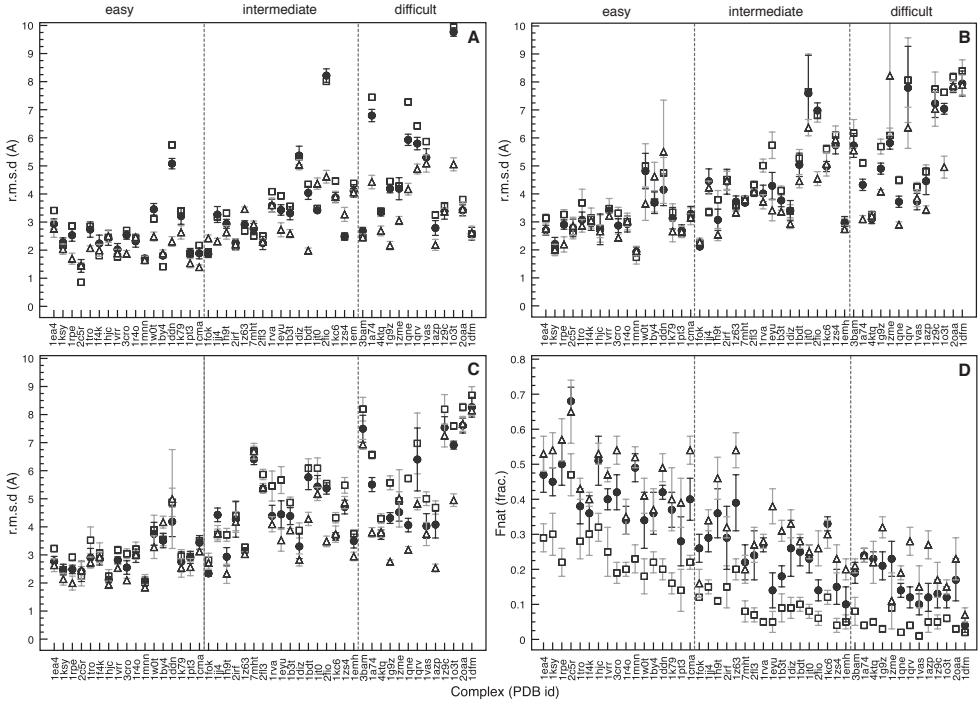



**Figure 6.4.** All heavy atom RMSD values from the reference complex (**A**: DNA, **B**: complex, **C**: interface) and fraction of native contacts (Fnat, **D**) for the 10 best solutions of the best cluster according to the HADDOCK score after rigid-body docking (open squares) and semi-flexible refinement (closed circles) starting from a canonical B-DNA structural model and after semi-flexible refinement (open triangle) starting from an ensemble of custom-built DNA models.

often large DNA conformational changes observed in the benchmark (ranging from 3 Å up to 10 Å).

The amount and consistency of the DNA conformational changes that did occur during semi-flexible refinement, can however provide an indication of the extend of conformational change to be expected in the final complex as we have shown before (326). By analyzing the conformational changes in the top ten solutions of the best cluster according to the HADDOCK score, we generated five new DNA structural models with custom conformations reflecting the conformational changes that took place in the DNA during the first docking round for every test case (see materials and methods).

The effects of using a custom-built DNA structural ensemble on the number of one and two-star solutions obtained is illustrated in Figure 6.3. The graph representation is similar to Figure 6.2, except that we now compare the docking results obtained after semi-flexible refinement starting from a canonical B-DNA structural model (Figure 6.2, final results) with those from a second docking run starting from the custom-built DNA structural ensemble. In a number of cases there is a marked increase in one and/or two-star solutions due to the use of the ensemble, while in other cases there is no improvement or even a reduction. However, because the ensemble contains custom built DNA structures in different conformations it is possible that one or several of these are less successful in sampling relevant

conformational space than the canonical B-DNA model used in the first run. However, if even only one of the five models is significantly better that canonical B-DNA, and the scoring and clustering stage select solutions obtained from this model then an improvement is achieved compared to only semi-flexible refinement. Figure 6.4 better illustrates the results by individual graphs showing for every test case the various RMSD values and fraction of native contacts for the 10 best solutions of the top-ranking cluster according to the HADDOCK score. The figure shows statistics for the corresponding solutions after semi-flexible refinement, the solutions from the rigid-body stage starting from canonical B-DNA, and the solutions after semi-flexible refinement using an ensemble of custom-built DNA starting structures. (Source data can be found in Tables 1 to 3 of the Appendix). For most complexes there is a marked improvement in terms of RMSD from the reference complex, when progressing from rigid-body docking to the use of an ensemble of custom built DNA structural models. The improvement in DNA, interface and all heavy-atom RMSD becomes more significant with the increasing difficulty of the test cases. This trend is to be expected as the conformational changes between unbound and bound structures are small in the "easy" category and become more pronounced in the "intermediate" and "difficult" categories of the benchmark. These results show the efficiency of the DNA modelling procedure in capturing the essential motions that occur in the DNA upon complex formation. The fraction of native contacts improves significantly throughout the benchmark even when the solutions improve little in terms of RMS deviations. Apart from this, the convergence in the 10 best solutions in general improves, which is apparent in the smaller standard deviations (Figure 6.4) and an improved clustering (Table 3, Appendix).

**Unbound-unbound docking using experimental derived restraints**

In a "real-life" docking situation, AIRs are typically defined based on experimental data or interface predictions (216,323). The quality and quantity of available data can influence the correct assembly of the interaction interface(s) and the conformational changes brought about in the flexible stages of the docking. To evaluate the performance of our two-stage protein-DNA docking protocol under these circumstances we selected six representative test cases from the "easy", "intermediate" and "difficult" categories of the benchmark (two of each). These are respectively the protein-DNA complexes formed by the phage 434 Cro (3cro) transcription factor and retinoid X receptor (1by4), the hyperthermophile chromosomal protein SAC7D (1azp) and papillomavirus type 18 E2 (1jj4) protein, the homing endonuclease I-PpoI (1a74) and the proline utilization transcription activator PUT3 (1zme). For these we defined AIRs based on experimental data collected from literature sources (see Material and Methods).

Docking the protein and DNA in their bound conformation (Table 5.1, bound-rigid) using rigid-body energy minimization only illustrates that the AIRs defined based on experimental data are also able to reconstruct the correct interaction interface(s) in all cases resulting in high quality predictions. The overall results for the unbound docking again show a significant improvement in terms of RMSD from the reference complexes and fraction of native contacts when progressing from rigid body docking to semi-flexible refinement and finally a second docking round starting from an ensemble of custom-built DNA structural models (Table 5.1). The best docking solutions superimposed onto their reference structures are presented in Figure 6.5.

Although the overall results improved for all six test cases, differences where observed. The bound and unbound components of the

**Table 6.1.** Performance of the two-stage docking protocol when using AIRs based on experimental information: average RMSD values from the target and fraction of native contacts for the top ten docking solutions of the top ranking cluster.

| | RMSD (Å) | | | | Fnat[e] | CAPRI[f] |
|---|---|---|---|---|---|---|
| | Total[a] | Interface[b] | DNA[c] | Protein[d] | | *,**,*** |
| **"easy"** | | | | | | |
| 1by4 | | | | | | |
| Bound rigid | $0.41_{0.08}$ | $0.34_{0.07}$ | $0.00_{0.00}$ | $0.38_{0.07}$ | $0.89_{0.02}$ | 0,0,10 |
| Unbound rigid | $4.33_{0.72}$ | $4.01_{0.53}$ | $1.41_{0.00}$ | $4.66_{0.73}$ | $0.11_{0.04}$ | 4,0,0 |
| Unbound flex | $6.72_{2.10}$ | $5.87_{1.71}$ | $1.90_{0.19}$ | $6.98_{2.21}$ | $0.17_{0.05}$ | 5,0,0 |
| DNA lib | $5.52_{2.43}$ | $4.91_{2.32}$ | $1.61_{0.14}$ | $5.85_{2.46}$ | $0.27_{0.09}$ | 4,3,0 |
| 3cro | | | | | | |
| Bound rigid | $0.32_{0.16}$ | $0.38_{0.19}$ | $0.00_{0.00}$ | $0.44_{0.22}$ | $0.85_{0.09}$ | 0,0,10 |
| Unbound rigid | $3.79_{0.60}$ | $3.51_{0.63}$ | $3.70_{0.00}$ | $3.50_{0.83}$ | $0.15_{0.05}$ | 10,0,0 |
| Unbound flex | $3.57_{0.63}$ | $3.29_{0.68}$ | $2.86_{0.30}$ | $3.19_{0.68}$ | $0.27_{0.07}$ | 6,2,0 |
| DNA lib | $2.89_{0.40}$ | $2.62_{0.73}$ | $2.08_{0.21}$ | $2.96_{0.43}$ | $0.40_{0.06}$ | 3,7,0 |
| **"intermediate"** | | | | | | |
| 1azp | | | | | | |
| Bound rigid | $0.33_{0.07}$ | $0.31_{0.07}$ | $0.00_{0.00}$ | $0.11_{0.00}$ | $0.92_{0.03}$ | 0,0,10 |
| Unbound rigid | $7.12_{2.06}$ | $7.09_{2.25}$ | $3.25_{0.00}$ | $3.58_{0.02}$ | $0.02_{0.02}$ | 0,0,0 |
| Unbound flex | $6.90_{2.00}$ | $6.68_{2.26}$ | $2.87_{0.32}$ | $3.64_{0.13}$ | $0.04_{0.04}$ | 0,0,0 |
| DNA lib | $4.56_{0.79}$ | $4.00_{0.45}$ | $1.83_{0.26}$ | $3.76_{0.16}$ | $0.10_{0.04}$ | 5,0,0 |
| 1jj4 | | | | | | |
| Bound rigid | $0.39_{0.10}$ | $0.40_{0.09}$ | $0.00_{0.00}$ | $0.10_{0.03}$ | $0.82_{0.07}$ | 0,0,10 |
| Unbound rigid | $4.23_{0.37}$ | $4.76_{0.48}$ | $3.19_{0.00}$ | $1.47_{0.05}$ | $0.09_{0.02}$ | 3,0,0 |
| Unbound flex | $4.25_{0.43}$ | $4.55_{0.58}$ | $3.19_{0.21}$ | $2.40_{0.02}$ | $0.16_{0.07}$ | 6,0,0 |
| DNA lib | $3.22_{0.30}$ | $3.62_{0.38}$ | $2.38_{0.14}$ | $2.37_{0.05}$ | $0.21_{0.07}$ | 9,1,0 |
| **"difficult"** | | | | | | |
| 1a74 | | | | | | |
| Bound rigid | $0.06_{0.01}$ | $0.07_{0.01}$ | $0.00_{0.00}$ | $0.01_{0.00}$ | $0.84_{0.01}$ | 0,0,10 |
| Unbound rigid | $5.43_{0.99}$ | $6.88_{0.97}$ | $7.44_{0.00}$ | $1.68_{0.14}$ | $0.04_{0.02}$ | 0,0,0 |
| Unbound flex | $4.95_{0.38}$ | $6.30_{0.46}$ | $7.12_{0.32}$ | $1.84_{0.14}$ | $0.14_{0.04}$ | 8,0,0 |
| DNA lib | $2.72_{0.25}$ | $3.37_{0.32}$ | $3.76_{0.19}$ | $1.78_{0.12}$ | $0.24_{0.05}$ | 9,1,0 |
| 1zme | | | | | | |
| Bound rigid | $0.48_{0.11}$ | $0.46_{0.08}$ | $0.00_{0.00}$ | $0.01_{0.00}$ | $0.79_{0.06}$ | 0,0,10 |
| Unbound rigid | $6.29_{0.64}$ | $5.49_{0.68}$ | $4.28_{0.00}$ | $5.67_{0.61}$ | $0.06_{0.03}$ | 0,0,0 |
| Unbound flex | $6.15_{0.62}$ | $5.29_{0.59}$ | $4.68_{0.33}$ | $5.88_{0.27}$ | $0.12_{0.06}$ | 4,0,0 |
| DNA lib | $5.27_{0.62}$ | $4.63_{0.80}$ | $3.35_{0.13}$ | $5.55_{0.48}$ | $0.15_{0.04}$ | 8,0,0 |

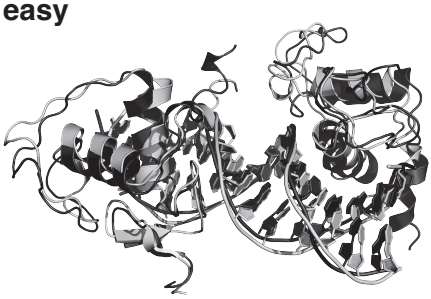
Average all heavy atom RMSD values from the reference structure (Å, standard deviation in subscript) calculated over the entire complex (a), the interface (b), the DNA (c) and the protein (d) for the ten top ranking solutions. The RMSD values are reported for; bound rigid-body docking (bound rigid); unbound rigid-body docking (unbound rigid), semi-flexible refinement (unbound flex.) starting from canonical B-DNA; unbound semi-flexible docking using a library of custom-built DNA structural models as input (DNA lib.). e) Fnat is the fraction of native contacts. f) number of one-, two- and three-star CAPRI ranked solutions obtained in the top ten solutions.

retinoid X receptor – DNA complex (1by4) differ little from each other in terms of RMSD from the reference and rigid body docking readily generates one-star solutions. The complex is composed of two proteins that interact with the DNA major groove but not with each other. Independent movement of both proteins resulted in a relative large variation in the ten best solutions after semi-flexible refinement when starting from a canonical B-DNA model. The use of a custom built DNA library does not reduce this variation but does significantly improve the fraction of native contacts and medium quality solutions. The phage 434 Cro – DNA complex (3cro) is a similar case with the exception that the proteins dimerize. This results in far less variation in the ten best solutions after the flexible stages and a sequential improvement of the RMSD values and fraction of native contacts at each step of the docking. The hyperthermophile chromosomal protein SAC7D – DNA complex (1azp) binds in a non-specific manner to the DNA minor groove. The experimental data available for this complex are less well defined than for the other test cases. Despite this, the two-stage docking protocol did reproduce the characteristic minor groove widening observed for this system resulting in a significant improvement in RMSD when using an ensemble of custom built DNA structural models. The specific kink in the DNA structure observed at the 2nd C-G base pair (61°) in the target complex was, however, predicted at the 3rd G-A base pair step (~25°) in the docking solutions. The potential of our two-stage docking protocol to deal with large DNA conformational changes is best illustrated in the case of the homing endonuclease I-PpoI – DNA complex (1a74). Here, the overall bend of ~38° is reproduced in the best solutions (~45°). The information available for this complex results in a well defined, curved, interaction interface on the protein and indicates that there is little conformational difference of the protein in its bound and unbound state. As such, the sharp bend introduced in the DNA by the analysis and modelling step could be sampled up to 10 times the standard deviation from the average to match the protein surface (see Materials and Methods). The proline utilization transcription activator PUT3 (1zme) is a difficult case from both protein and DNA perspectives. The protein contains two globular DNA binding domains connected to a core domain with a long flexible linker. The NMR ensemble of the unbound protein contains the DNA binding domains in many different orientations that prevent effective docking in the rigid body stage. Therefore, we cut the protein at the flexible linkers, resulting in three parts that were docked as separated bodies. Peptide linker restraints were defined between de amino acids at the scission sites. After semi-flexible refinement we reconnected the different parts in the ten best solutions and used the resulting protein ensemble for the second docking stage starting from an ensemble of custom built DNA structural models.

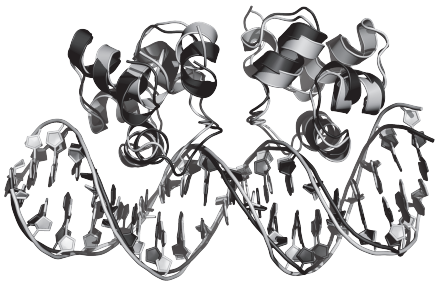
## Discussion

The use of Ambiguous Interaction Restraints (AIRs) is essential to the success of the HADDOCK docking methodology in general. These are used to position the protein at the interaction interface of the DNA and, together with the flexible stages of the docking, to facilitate conformational changes. We have shown previously the importance of AIRs in protein-DNA docking (326) using three monomeric transcription factor DNA complexes as test cases. In the current study we refined our initial method and evaluated its performance on a benchmark of 47 protein-DNA complexes (327). Compared to the initial three test cases the benchmark contains complexes from various structural functional classes in which one or multiple proteins interact with the DNA using various binding modes. Because of the presence of multiple proteins or DNA-binding domains,

**easy**



1by4** 0.40[a] 3.55[b] 1.50[c]



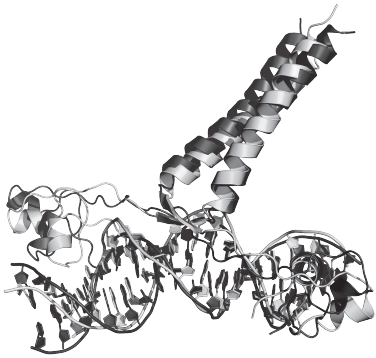
3cro** 0.50[a] 2.23[b] 1.93[c]

**intermediate**

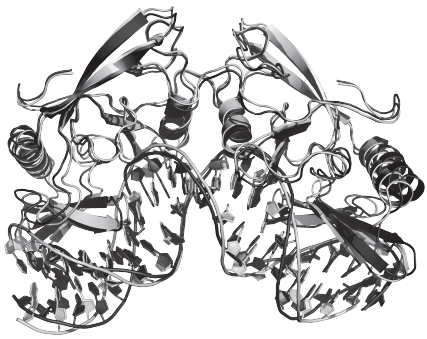
1azp* 0.11[a] 3.44[b] 1.58[c]

1jj4** 0.44[a] 2.63[b] 2.26[c]

**difficult**



1zme* 0.15[a] 3.75[b] 3.23[c]



1a74** 0.31[a] 3.24[b] 3.70[c]

***Figure 6.5.** Best solutions from unbound flexible docking using an ensemble of custom-built DNA structural models (blue) superimposed on to the reference structure (yellow). The complexes are grouped according to their docking difficulty ("easy", "intermediate" and "difficult) as indicated in the benchmark. De CAPRI score for each solution is indicated as one or two stars after the PDB code as well as the fraction of native contacts (a), the interface (b) and DNA RMSD (c) from the reference structure. RMSD values (Å) were calculated after superimposition on all heavy atoms of the selected regions. The figures were generated using Pymol (DeLano Scientific LLC, www.pymol.org).

40% of the benchmark required docking following a multi-body (N>2) approach. This challenging benchmark offers a good platform to evaluate the capabilities of our docking method. We will discuss in the following the two questions that were the focus of both this study as well as the previous work describing the initial protein-DNA docking method.

**How well is the method able to identify the correct interaction interface(s)?**

The assembly of the interaction interface(s) is a process driven by AIRs. In "real-life" docking settings the AIRs are typically defined based on experimental data and/or interface predictions. The quality of the docking solutions is therefore closely related to the amount and quality of available data in terms of their accuracy and information content. We started from an ideal situation in which the restraints were derived from the intermolecular contacts in the reference complex. Bound docking resulted, in nearly all cases, in high quality (three stars) predictions among the top 10 solutions based on the HADDOCK score (Figure 6.1). The percentage of generated high quality solutions, however, declined for the 'intermediate' and 'difficult' cases indicating that interface topology features such as segmentation and interpenetration of structure elements are challenging to model. Scaling down of the non-bonded interactions in the rigid-body docking stage was successfully applied to a number of such cases where interpenetration of the interfaces during docking was necessary to assemble the final complex. The unbound two-stage flexible docking using the same restraints (Figure 6.2-6.4) resulted in the prediction of acceptable to medium quality solutions depending on the level of difficulty of the test cases. Although these results are significantly better than unbound rigid-body docking only, they still indicate that conformational changes are the limiting factor in protein-DNA docking.

The same series of docking experiments were performed with a representative selection of six test cases using AIRs defined based on experimental information (Table 6.1, Figure 6.5). The results were comparable to the use of ideal restraints in terms of the CAPRI quality criteria. This clearly illustrates that readily available non-structural experimental data are sufficient to assemble the correct interaction interface(s) in these challenging, often multi-component, protein-DNA systems. Still, the quality of the generated solutions is directly related to the quality of the used experimental information. Sparse and/or low quality information will likely result in poor quality docking solutions, especially for multi-component systems. The Ambiguous Interaction Restraints can, however, be defined based on a wider variety of information sources than used in the current work. For instance, NMR data or even statistical protein-DNA interaction potentials, are promising means of improving the results either by driving the docking or filtering solutions afterwards. With respect to the latter we should note that the many different solutions generated in this benchmark docking effort, provide a compelling set of decoy structures that can be useful for the development and validation of scoring functions.

**How successful is the method in dealing with conformational changes upon complex formation?**

The correct treatment of conformational changes upon complex formation is likely the most challenging aspect of protein-DNA docking. Both protein(s) and DNA readily change their conformation upon complex formation. The extend of this change forms the basis of the protein-DNA benchmark categorization. Our two-stage protein-DNA docking method was designed to deal with this challenge and its performance is best illustrated in the docking of unbound proteins with canonical B-DNA using ideal AIRs. While a single docking run was sufficient to

generate medium quality solutions for the "easy" cases, the two-stage protocol was often required to generate acceptable to medium quality solution for the "intermediate" and "difficult" cases. This approach was successful in generating acceptable solutions for 94% of the complete benchmark. This illustrates that the explicit flexibility implemented in HADDOCK is sufficient to generate acceptable or higher solution in the "easy" cases where conformational changes are limited but that this approach fails for cases where such changes are more pronounced such as in the "intermediate" and "difficult" cases. For the latter, our DNA analysis and modelling procedure is capable of extracting the main bend and twist motions that occur in the DNA upon complex formation and use these for the benefit of DNA modelling. In that way, a larger part of the relevant DNA conformational space can be sampled than what is feasible within a single round of semi-flexible refinement. Even results of the "easy" test cases with limited conformational changes are improved by this two-stage procedure. Finally, the use of experimentally-derived AIRs on a subset of six test cases showed that our method also significantly improved the docking results under real-life conditions when less ideal AIR restraints are available.

Although the semi-flexible refinement stage of HADDOCK is able to introduce many of the DNA conformational changes required for correct complex formation it has difficulties predicting DNA groove expansion facilitated by negative base pair step sliding (for example in 1a74 and 1g9z). Consequently, this mode of conformational change is not detected by our DNA analysis procedure and not introduced in the custom-built DNA ensemble. Although the improvements in RMSD to the reference complex and fraction of native contacts clearly illustrate that our method outperforms rigid-body docking it does raise questions on the quality of the DNA in the generated solutions. This however, remains a difficult issue due to the lack of DNA structure validation procedures. Furthermore, our method predominantly focuses on the conformational changes in the DNA, but also proteins can often change their conformation upon complex formation, sometimes quite drastically as, for example, in the restriction endonuclease MVAI (2oaa). While accounting for small conformational changes by means of flexible refinement and the use of protein ensembles that sample different interface conformations, large conformational changes such as loop and domain rearrangements or disordered to order transitions remains a challenge. Such events are present in some of the test cases where the use of an ensemble of custom-built DNA structural models did not improve the results significantly. This still leaves plenty of opportunities for improvements, for instance in those cases where protein domain rearrangements are facilitated by flexible "hinges" connecting them. Such domains can be docked as separate bodies, enabling them to sample conformational space individually. This procedure has been successfully used for the proline utilization transcription activator PUT3 (1zme) in this study.

The flexible protein-DNA docking approach described in this paper can benefit protein-DNA interaction studies at several levels. It can be used to generate models of protein-DNA complexes from the structures of the unbound proteins and a canonical B-DNA in the presence of suitable experimental data without any prior knowledge of the DNA conformational changes required to establish the complex. It should also be useful for studying the effects of mutations or different operator sequences on complex formation. In addition, it can assist in experimental structural studies by, for instance, providing initial DNA structural models to guide and speed up the NMR analysis and assignment process.

In summary, by allowing the inclusion

of a large variety of experimental and/or prediction data, together with a flexible description of the DNA, the proposed docking approach should be a useful tool in structural studies of protein-DNA complexes.

## Materials and Methods

### Protein-DNA docking benchmark

The performance of HADDOCK was evaluated using the coordinate files for the bound and unbound proteins of 47 protein-DNA complexes available in the protein-DNA benchmark version 1.2 (http://haddock.chem.uu.nl/dna/benchmark.html (327)). Canonical B-DNA 3D structural models were built using the 3D-DART web server (http://haddock.chem.uu.nl/dna (328)). Their conformation was of BII type with the sugar pucker in the C2'-endo conformation (sugar pseudo-rotation phase angle (P) = 155°, DNA backbone torsion angles: $\alpha$ = 309°, $\beta$ = 159°, $\gamma$ = 37°, $\delta$ = 146°, $\varepsilon$ = 218°, $\zeta$ = 191° and $\chi$ = 260°).

### Restraints used in the docking

*Ambiguous Interaction Restraints (AIR), based on the true interface:* Ideal AIR restraint sets were generated based on the true interface(s) of the reference complexes as follows; 1) Retrieval of all intermolecular atom-atom contacts below a cutoff of 5.0 Å. Contacts that originated from amino-acid residues having a relative main-chain or side-chain solvent accessibility of less than 30% as measured by NACCESS (134) where discarded.; 2) Transformation of the atom-atom contacts to their respective residue-residue counterparts distinguishing between three categories: amino-acid to nucleotide base contacts, amino-acid to nucleotide sugar-phosphate backbone contacts or amino-acid to full nucleotide contacts.

All residues used in creating the interaction restraint file were defined as 'active'. In effect we used the same procedure to generate AIRs as in the case of experimental information with the difference that they are only defined between the residues that are known to be in close vicinity in the reference complex.

*Ambiguous Interaction Restraints (AIR), based on experimentally information:* To evaluate the performance of HADDOCK in docking protein-DNA complexes using experimental information we selected six representative tested cases from the "easy" (3cro, 1by4), "intermediate" (1azp, 1jj4) and "difficult" (1a74, 1zme) category of the benchmark. For these we collected biochemical and biophysical information from literature sources. Only residues that are solvent accessible in the unbound proteins, using the same criteria as described above, were considered. For those DNA bases shown to be involved in specific interactions with the protein, only atoms able to interact by hydrogen-bond or non-bonded interactions were defined. This selection was further subdivided into atoms facing either the major or minor groove in case information about the protein-binding mode was available (Table 6.2). In case of non-specific interactions with the DNA only the atoms of

**Table 6.2.** Nucleotide atom subsets used in the definition of AIRs.

| DNA base | Minor groove atoms | Major groove atoms |
|---|---|---|
| Thy | H3, O2, C2' | H3, O4, C4, C5, C6, C7' |
| Ade | N1, N3, C2, C4' | H61, H62, N1, N7, C5, C6, C8' |
| Gua | H1, H21, H22, N3, C2, C4' | H1, H21, N7, O6, C5, C6, C8' |
| Cyt | N3, O2, C2' | H41, H42, N3, C4, C5, C6' |
| | **None-specific backbone atoms** | |
| Sugar-phosphate backbone | C1', C2', O3', O5', P, O1P, O2P | |

Individual subsets are defined for those atoms facing the DNA major groove and minor groove for the four bases and for the sugar-phosphate backbone atoms.

the sugar-phosphate backbone that are able to interact via hydrogen bonds or non-bonded interactions were defined (Table 6.2).

Solvent accessible residues located in the predicted interaction interface, for which no experimental information was available, were defined as passive. Residues for which experimental information was available were defined as active. An overview of the data used is listed in Table 6.3.

*DNA restraints:* In order to preserve the helical conformation during the flexible stages of the docking the DNA was restrained as described before (326). For the docking of the unbound protein(s) to a canonical B-DNA structural model, the dihedral angles of the sugar-

**Table 6.3.** Definition of the AIRs based on experimental data for the six selected test cases.

| | Protein | DNA | Reference |
|---|---|---|---|
| **"easy"** | | | |
| 1by4 (357) | Act: (K31,R32)[ab] → T5,C6,G25,A26 (E24,K27)[ab] → G3/4,C27/28 (K72,K73,R80)[b] → A2,G3/4 Pas: V34,A75,V76,Q77, R55,N56,Q59,R62 | Act: (T5,C6,A26,C27,C28,T29)[a] (G3,G4)[acd], (A2,T24)[ac] T23[c],G25[ad] | (67,114,126, 161,182,202, 231,256,318) |
| 3cro (222) | Act: (K29,Q31,S32,K42-P44)[a] L35[b] → C14,T15/T23,33 Pas: K9,T18-T20,G27,V28,Q30,Q34, I36,E37,V40,T41,R45,F46 | Act: (C6,A7,T16-T18,C24,A25,T34-T36)[a],(T32,T33)[abc] (T4,A5,T13,C14,T15,T22,A23, G31)[ac] | (125,164,165, 345) |
| **"intermediate"** | | | |
| 1azp (357) | Act: W24[e] → G3,G15 V26[b],M29[b],S31[e],V45[e] → C2-A4, T13-G15 (K22,T33,R42)[e] → T5-G7,C10-A12 Pas: K21,R25,G27,K28,K39,T40,A44, S46,E47 | Act: C2,G3[f],A4,T5,C6,G7, C10,G11,A12,T13,C14,G15[f] | (61,87,147, 251) |
| 1jj4 (357) | Act: (N13,K16,C17,R19-R21)[a] Pas: S34,T35,H37 → T26-C27 | Act: (A3,C4,T30)[a] ,(C5,G28,G29)[ad] (T25-C27)[c] | (23,268) |
| **"difficult"** | | | |
| 1a74 (357) | Act: (H97,N122)[ab] → A35,G36 (A54-N56,T59,R60,R65,R73, G75)[a] → T1-C7 Pas: V51,G57,P58,T66,V71,H77, H100,K119 | Act: (T1-C7)[abd] ,(A35,G36)[b], G40[d] | (14,92,95, 109,227,348) |
| 1zme (357) | Act: (R9,R11,H12,R80,R82,H83)[a] Pas: A4,K14,K39-S43, A75,K85,K100-S114 | Act: (C2,G3,G4,C15,C17,G18, C20,G21,G22,C33,C34,G35)[a] (T26-C32,C9-T14) | (17,39,208, 209,292,337) |

Active residues (Act) are grouped according to the available information. Continues stretches of residues are separated by a dash. Arrows indicate active restraints for specific pairs of residues. Passive residues (Pas) are only defined for the protein. Since 1by4, 1jj4 and 1a74 are symmetrical dimers only the restraints for one subunit are shown. Base-specific restraints for 3cro, 1by4, 1jj4, 1a74 and 1zme are targeted to the atoms of the nucleotides facing the major groove and those of 1azp to those facing the minor groove (Table 6.2). [a]Conserved residues; [b]Mutagenesis data; [c]Ethylation interference data; [d]Methylation interference data; [e]NMR native state amide hydrogen exchange; [f]Raman spectroscopy.

phosphate backbone of the input structure (inp) were measured and used as restraints (restricted to $\alpha = \alpha_{inp} \pm 10°$, $\beta = \beta_{inp} \pm 40°$, $\gamma = \gamma_{inp} \pm 20°$, $\delta = \delta_{inp} \pm 50°$, $\varepsilon = \varepsilon_{inp} \pm 10°$ and $\zeta = \zeta_{inp} \pm 50°$). For the docking of the unbound protein(s) to the ensemble of custom-build DNA structural models the restraint were defined in the same way but with error values were reduced to half of those in the canonical B-DNA case.

**Docking protocol**

The default protein-DNA docking protocol (326) implemented in HADDOCK version 2.1 (74) was used for all the docking runs. Several docking-specific modifications were made:

Bound-bound docking: only rigid body docking generating 2000 solutions. Non-bonded interactions were scaled down during the docking for those cases in which complex formation was prevented due to non-bonded repulsions associated with interface penetration during the rotation and translation stages. The scaling factor (inter_rigid in the run.cns file) was set to 0.01 (* Fig. 6.1-6.3) or 0.00001 (•, Fig. 6.1-6.3)

Unbound-unbound docking using a canonical B-DNA structural model: A single component HADDOCK run was performed using the unbound proteins to yield a better sampling of side-chains and loop conformations. The residues of the interface (either defined based on the reference complex or on experimental information) were allowed to sample additional conformations during the semi-flexible refinement stage. Here, semi-flexible refinement signifies the combination of the semi-flexible simulated annealing stage in torsion angle space and the final water refinement stage in Cartesian space. Four solutions and the original unbound structure were together used as an input ensemble for unbound-unbound docking. A total of 4000 docking solutions were generated in the rigid body docking stage and the top 10% based on the HADDOCK score were used in the subsequent semi-flexible refinement stage. During the semi-flexible simulated annealing stage, the full DNA excluding the terminal base pairs was treated as semi-flexible. The amino-acid residues within 5.0 Å of any partner molecule were automatically defined as semi-flexible.

Unbound-unbound docking using custom-build DNA structural models: The same protocol as for unbound-unbound docking starting from canonical B-DNA was used with as only difference that the conformational freedom of the DNA in the semi-flexible simulated annealing stage was limited by automatically defining both the amino-acid residues and nucleotides within 5.0 Å of any partner molecule as semi-flexible and by reducing the error range by half for the sugar-phosphate backbone dihedral angles as described above. The procedure for generating custom DNA structural models used as input for this docking run is described below.

**Generation of custom DNA structural models**

The generation of custom DNA structural models is based on an analysis and a modelling step.

Analysis: The ten best solutions from the top ranking cluster according to the HADDOCK score were selected. The DNA structures in these solutions were analyzed using 3DNA (194,195) and the DNA bend analysis algorithm implemented in the 3D-DART server (328). This resulted in average parameter values for the six base pair (step) parameters (79) for every base pair (step) in the structure. These describe the conformation of the DNA. The average global bend vector with respect to a common reference frame between every successive base pair in the structures was calculated by 3D-DART. This information was used in the modelling stage.

Modelling: the modelling of custom DNA structures is based on the progressive introduction of global and local DNA conformational changes to a canonical B-DNA starting model;

1. A default set of base pair (step) parameters

representing a canonical B-DNA conformation with the same sequence as the target structure is generated by 3D-DART using the "fiber" utility of the 3DNA software suite.

2. The Roll and Tilt values in the default set are updated by 3D-DART to reflect the average global bend vector for every base pair step in the sequence. The central base pair is used as origin of the global reference frame and default Twist values are used for correcting the vectors direction relative to the reference frame. The introduced bend vector between base pairs is scaled, enabling sampling of conformation change beyond the limits of the values defined by the average ± the standard deviation determined in the analysis stage. The scaling factor is set between 2.0 and 3.0 for those ensembles that show little deviation from a canonical helix and between 4.0 and 6.0 for the remaining test cases. For the docking of 1a74 using experimentally derived restraints the scaling factor was set to 10.0 to match the amount of DNA bend to the curved interaction surface of the protein (see Results).
3. All base pair step parameters are updated to reflect the average values as determined by the analysis stage resulting in a new weighted parameter $P_{Wxi}$ at base pair step i defined as follows:

$$P_{Wxi} = \left(2 - \left(\sqrt{\sigma_{pi}/\sigma_{\Sigma p}}\right)^{S}\right) * P_{xi} \quad (6.2)$$

Were $P_{xi}$ is the average value for the given parameter at base pair step i obtained from the analysis stage, $\sigma_{pi}$ defines the standard deviation for the given parameter at base pair step i and $\sigma_{\Sigma p}$ the standard deviation for the given parameter for all base pair steps. S is a parameter-specific scaling factor that compensates for the over- or underestimation of a given parameter as a result of the HADDOCK semi-flexible refinement stages. S was set to; Twist: 0.8, Roll: 0.8, Tilt: 0.8, Rise: 0.0, Slide: 0.2 and Shift: 0.8.

The new value $P_{ni}$ for the parameter at base pair step i is now calculated as follows:

$$P_{ni} = P_d + \left((P_{Wxi} - P_d) * V\right) \text{(Eq. 6.3)}$$

Here $P_d$ is the default value from canonical B-DNA for the given parameter at base pair step i and V is a variance value used to sample the parameter above or below its adjusted average (set to 0.8 by default).

4. The default base pair parameters are updated in the same way as for the base pair step parameters. The base pair parameter-specific scaling factors (S) used are: Shear: 1.0, Stretch: 1.0, Stagger: 1.0, Buckle: -1.0 and Propeller Twist: -1.0. The variance parameter V is set to 0.8 by default.
5. The updated list of base pair (step) parameters is used to build a 3D DNA structure using the same parameters for the sugar pucker and phosphate backbone dihedral angles as in the case of canonical B-DNA.

**Analysis**

The quality of the generated solutions was evaluated using the CAPRI criteria expressed as stars; "high" (3 stars): Fnat > 0.5, l-RMSD or i-RMSD < 1.0 Å; "medium" (2 stars): Fnat > 0.3, l-RMSD < 5.0 Å or i-RMSD < 2.0 Å; "acceptable" (1 star): Fnat > 0.1, l-RMS < 10.0 Å or i-RMSD < 4.0 Å. Fnat is the fraction of native contacts within a 5 Å cutoff, l-RMSD is the ligand backbone RMSD from the target, after superimposition on the receptor, and i-RMSD the interface residues backbone RMSD from the target. The RMSD values were calculated using ProFit (A.C.R. Martin, www.bioinf.org.uk/software/profit) using all Cα

atoms for the protein and all phosphate atoms for the DNA.

**Hardware**

HADDOCK docking runs were performed on a Transtec (Transtec AG, Tubingen, Germany) computer cluster operating with 48, 2.0 GHz, 64 bit Opteron processors. As a measure of CPU requirements, one complete run starting with 4000 structures in the rigid-body docking stage could be performed in 4 h on 48 processors.

# Conclusions and Perspectives

Chapter 7

The protein-DNA docking field has seen a progressive development over the past decade, resembling that of the protein-protein docking field. The tests systems used in the early years of method development consisted of small transcription factor - DNA complexes which were docked in a rigid fashion often with the protein and/ or the DNA in their bound conformation. In addition, many of these systems were docked in their monomeric state instead of their biological relevant dimeric state. The emphasis was on the reconstruction of the interaction interface between both molecules. Gradually over time, test systems became more complex and molecular flexibility in complex formation started to be considered. This meant that the search through conformational space in order to assemble the correct interaction interface became more demanding as a result of the additional number of degrees of freedom that need to be included to account for molecular flexibility. These developments were often triggered by new algorithms and computational techniques pioneered in the protein-protein docking field.

The two docking fields, however, did not fully develop in parallel. New developments in protein-DNA docking are often fragmented: they often result of innovations in a protein-protein docking approach that have been extended to deal with protein-DNA systems. In contrast to this, developments in the field of protein-protein docking of the last few years have very much been the result of a community effort. This has been driven by the use of common standards for evaluating the performance of docking approaches and the availability of a protein-protein docking benchmark. Furthermore, the various research groups can participate in CAPRI (Critical Assessment of Predicted Interactions, http://capri.ebi.ac.uk, (138)) a community wide "blind docking experiment". This is the ultimate test were docking methods can be put to the test on not yet published complexes. Since the start of the CAPRI experiment in 2001 there have been 41 docking targets. This community-wide effort has had a dramatic effect on the development curve of the protein-protein docking methodology.

Up to now, protein-DNA systems have not been a part of CAPRI with exception of one protein-RNA systems (target 33) for which HADDOCK made a successful prediction. This does not mean that protein-DNA systems cannot be a part of the CAPRI competition. The limitation here is the availability of such targets and the willingness of researchers to provide their data ahead of publication. The initial purpose of CAPRI is to serve as an unbiased test for evaluating the performance of a method; it thus not primarily a mean for method development. CAPRI often requires the use of a homology modelling approach to generate the unbound structure for one of the molecules that need to be docked. As such, the quality of the docking predictions not only depends on the accuracy of the docking method but also on the accuracy of the homology modelling procedure. It is therefore important that the method developers have access to a rich, standardized benchmark that can be used for method development prior to participation in CAPRI. Such a benchmark will also provide a scale with which the docking difficulty of a CAPRI target can be assessed.

For the protein-protein docking field such a benchmark has been constructed and is widely used (56,220). In chapter 5 of this thesis a similar dedicated benchmark for protein-DNA systems has been described. With a collection of 47 diverse protein-DNA complexes the benchmark contains enough challenges to serve the docking community for some years to come. With the discussion of the initial protein-DNA docking method

in chapter 3 and the large-scale benchmark docking effort in chapter 6 the HADDOCK two-stage protein-DNA docking method has been validated. This lays the ground for further developments and will hopefully stimulate other research groups to test their methods in a similar manner using our benchmark.

## Perspectives on the performance of the two-stage protein-DNA docking method

The protein-DNA benchmark was initially categorized into three groups based on system characteristics such as the number of individual components and the conformational changes that unbound components undergo upon complex formation. The docking results (chapter 6) described in terms of CAPRI scores correspond well with this classification. This indicates that, in case of HADDOCK, there is a good correspondence between the docking performance and the predicted difficulty, expressed as "easy", "intermediate" or "difficult", in the initial classification scheme of the benchmark. A similar relationship has also been found for protein-protein docking (321). The ranking of all test cases within the three categories is, however, likely to differ between methods as they all have their own weaknesses and strong points. With respect to this, the performance of HADDOCK on the benchmark scale can be assessed together with perspectives on possible feature developments.

The "easy" category of the benchmark contains complexes that have been commonly used in the last decade for method development and validation such as the many small transcription factor-DNA complexes. Using a two-stage protein-DNA docking approach, HADDOCK readily predicts medium quality solutions (CAPRI two stars) when experimental data are used to drive the docking; three-star predictions are within our reach when ideal restraints are used. In contrast to this, rigid-body docking only stage results most often in one-star and only few two-star predictions (chapters 3 and 6). This indicates that, although conformational changes upon complex formation are relatively small in the "easy" category, they are nevertheless vital; explicitly considering these in the docking procedure improves the results significantly. As such, methods that incorporate molecular flexibility by implicit and/or explicit means outperform methods that do not account for any flexibility; they are likely to be the only ones able to generate two- or three-star predictions in a CAPRI setting. It is striking to see that HADDOCK is not able to generate three-star predictions even when true interface restraints are used (chapter 6). This indicates that the treatment of flexibility is still a limiting factor. This plateau in HADDOCK's current performance might come from limitations in the force field and/or from the rather short molecular dynamics refinement stage that are unable to induce the subtle protein and DNA conformational changes required to generate three-star predictions. Refining the top scoring HADDOCK predictions in an additional molecular dynamics simulation could potentially solve this. The combined use of multiple methods is a reoccurring theme in today's docking and a means to compensate for the shortcomings of the individual methods. With respect to this, molecular dynamics simulations can also aid in aspects of complex formation for which docking is not the best method to provide answers. These are for instance the delicate and dynamic process of DNA recognition by amino acids or a more detailed description of the DNA conformation in the complex.

When shifting from the "easy" complexes to the "intermediate" and "difficult" ones one can observe a gradual but significant increase in the complexity of the protein-DNA interfaces and of the conformational

changes needed to establish the final complex. In this context, the ability of a docking method to establish the correct encounter complex and to facilitate the conformational changes required to form the final interaction interface becomes essential for the generation of acceptable solutions (CAPRI one star). In case of the "intermediate" test cases, HADDOCK is able to generate one-star solutions using rigid-body docking, both with true interface- as well as experimental derived restraints. Introducing flexibility by means of the semi-flexible refinement stage is required in order to generate two-star predictions.

For the "difficult" cases incorporation of flexibility is essential if any acceptable predictions are to be generated. For these complexes, the benefits of our two-stage docking approach are evident as the docking results often significantly improve (although not always). The cases were improvements are observed correspond to those complexes in which the DNA readily changes conformation, sometimes in dramatic ways. The "intermediate" and "difficult" categories, however, also contain quite a number of complexes in which also the protein(s) change(s) conformation upon complex formation, sometimes also in dramatic ways. Here, the benefits of our two-stage docking method are less apparent as the geometrical principles that allow for the modelling of DNA unfortunately do not apply to proteins. This does not mean that protein flexibility is not taken into account: aspects like side-chain rearrangements are easily accounted for, loop rearrangements are sampled by means of an ensemble representation and rigid domains separated by flexible hinges are docked separately (e.g. 1zme, chapter 6). Still there are a number of major conformational changes that cannot be dealt with at the moment. These are, for instance, large flexible domain motions or internal domain rearrangements (chapter 2). It is expected that these challenges will dominated the field for some time to come.

Establishing the (near) correct encounter complex seems to be required in the above-mentioned cases if the introduction of any conformational changes afterwards is to lead to any improvement. The information-driven docking approach in HADDOCK has proven to be very successful for these cases. In contrast, most other methods are in essence ab initio methods using the physicochemical characteristics of the molecules to guide the search. They do, however, often use additional information to either restrict the search space or filter the solutions. With the omnipresence of conformational changes in DNA and the regularity of its helix these methods are currently only able to deal with the "easy" test cases. On the other hand, the successful use of information to drive the docking is equally well HADDOCK's weakness. Without such information, HADDOCK performs poorly compared to ab initio methods such as ZDOCK and FTDOCK, at least for the "easy" cases. The efficiency of HADDOCK is therefore closely related to amount and quality of available data in terms of their accuracy and information content. The easy way in which HADDOCK can make use of additional information sources for the benefit of docking does provide ample opportunity for further developments. The use of statistical protein-DNA pairing potentials could be a promising addition in the search for the correct interaction interface. Similarly, additional information could be used for assessing the docking solutions in the final selection process. It is also important to realise that docking in general does not have to be the final stage in a research workflow. A docking method, in particular HADDOCK, can be an integral part of the workflow in which initial docking models can be used for experimental design and validation afterwards.

It should be clear from this discussion, that, at least for HADDOCK, a correct treatment of flexibility remains one of the most important challenges. The use of additional information sources, as described above, can help in the formation of the correct encounter complex, but the contribution to conformational adaptation is limited. As long as conformational changes in protein and/or DNA are restricted to side-chain and limited backbone rearrangements, HADDOCK is able to handle them in one docking run as illustrated for the "easy" test cases. Without an adequate number of restraints, however, the ability of modelling large side-chain and backbone rearrangements by HADDOCK is limited. This is true for many docking methods that incorporate explicit flexibility. These limitations impose a barrier for the exploration of a greater conformational space.

Current docking methods, including HADDOCK, try to lower this barrier by combining explicit with implicit means of treating flexibility such as the use of multiple starting models or additional refinement steps, e.g. using normal modes. The two-stage method described in this thesis follows a similar strategy. The assumption is that HADDOCK is able to initiate in the first docking run the conformational changes in the DNA that can lead to the final conformation in the complex. The subsequent DNA analysis and modelling step captures these changes and uses them to generate an ensemble of DNA models more closely reassembling the bound form. These are subsequently used in a second docking run. As such, the method does not require a particular starting DNA conformation for docking other than canonical DNA. The success is directly related to the ability of HADDOCK to introduce the specific conformational changes in the DNA and our knowledge of DNA structural dynamics to generate new custom-built DNA models. With respect to this, the best performance is obtained in cases where the global conformational changes are predominantly composed of bending and slight changes in the groove width. HADDOCK, however, has difficulties in expanding the DNA groove as a result of negative base pair step sliding. Consequently, this mode of conformational change is not detected by our DNA analysis procedure and not introduced in the custom-built DNA ensemble. Ideally this problem should be solved by a force field that more accurately describes base pair stacking. Alternatively, the DNA modelling step could be improved to detect such events and correct for them accordingly.

Finally the importance of water molecules at the interface of biomolecular complexes has been recognized for quite some time (49). So far interfacial waters in docking have been considered in only isolated cases for protein-protein (325), nucleic-acid ligand complexes (221) and protein-DNA systems (8). In all these cases, however, the inclusion of water did improve the docking results significantly. With the various functions proposed for water molecules at protein-DNA interfaces (chapter 2) a proper consideration of water molecules in the docking is expected to improve the results significantly.

## Perspectives on CAPRI for protein-DNA docking

We have discussed above the strengths and weaknesses of our two-stage protein-DNA docking approach. A direct comparison with other docking methods is however difficult to make. HADDOCK is the first method that can readily deal with large conformational changes by simultaneously docking more than two components. In addition, while we have consistently used CAPRI scoring criteria as means of validation, other methods have used a variety of criteria, making a comparison of the performance difficult. A difference in used criteria is in a sense understandable especially for the

intermolecular contacts. In protein-DNA complexes, in contrast to protein-protein complexes, there is a clear difference between stabilizing contacts, primarily involving the sugar-phosphate backbone, and functionally active contacts involving the bases. Protein-DNA interactions often show a level of plasticity in which the protein may interact with the DNA sugar phosphate backbone in number of slightly different poses while maintaining the functionally active contacts. This has important biological implications, especially for transcription factors, as described in chapter 2. The stabilizing contacts account for the majority of the intermolecular contacts made in an average protein-DNA complex. However, in calculating the fraction of native contacts, used in the CAPRI score, we refer to one reference complex with a fixed set of stabilizing sugar-phosphate backbone contacts. Even if the modelled interface predictions are of high quality the variations in DNA bending and twisting and their effects on the distribution and dimension of DNA grooves can alter the topology of the complex with respect to the target. This is likely to change the stabilizing contacts made and, therefore, negatively influence the CAPRI score.

As a consequence, one can argue if the current CAPRI scoring scheme is sufficient as quality measure for protein-DNA docking efficiency, reflecting the dynamical behaviour of these systems. This, at present, makes it difficult to know which solutions, if any, are correct.

## Perspectives on the practical use of the methods described in this thesis

Docking can be regarded as an academic exercise, a means of testing our current understanding of the principles underlying biomolecular behaviour and complex formation. In principle this is what docking is all about, the better we understand the systems we are studying the better we should be able to model them. Although this will remain an important aspect of docking it has now outgrown the theoretical phase and set to practical use. Nowadays it is clear that even the most efficient docking method cannot be regarded as successful if it is not put to practical use. Here, HADDOCK is doing very well, maintaining a close relationship with a large user community. This relationship between the developers



**Table 7.1.** Studies in which the HADDOCK, protein-DNA docking protocol was used. The information sources used to construct Ambiguous Interaction Restraints are listed.

| HADDOCK | Information sources used for AIR definition | Reference |
|---|---|---|
| CalC | CSP | (294) |
| Mrf2 | CSP, PR | (44) |
| Phr | EC | (189) |
| MyT1 | CSP, MU | (110) |
| THAP1 | CSP, MU | (31) |
| tvMyb1 | CSP, MU, RDC | (193) |
| LINE1-endonuclease | MU, EC | (259) |
| Endonuclease V | EC, IP | (203) |
| C4b | CSP, MU | (235) |
| 3 monomeric repressors | MU, EthI, EC | (326) |

Used abbreviations: CSP; Chemical Shift Perturbation data, EthI; Ethylation Interference, EC; Evolutionary conservation, PR; Paramagnetic spinrelaxation, MU; mutagenesis data, RDC; Residual Dipolar Coupling, IP; Interface Prediction

and the users has been a powerful drive in the development of the method. This is also true for the two-stage protein-DNA docking method described in this thesis. Ever since the publication of the initial method described in chapter 3 it has been successfully applied by a number of laboratories worldwide (Table 7.1).

Protein-DNA docking models of a number of studies, in which NMR data were used to drive the docking, have been deposited as experimental structures in the RCSB protein databank. To further stimulate the wide usage of a method, it is also important that its functionality becomes available to the user in an intuitive manner. Since the launch of the HADDOCK web server (http://haddock.chem.uu.nl; http://www.haddocking.org) in September 2008 the HADDOCK user community has grown fast. The intuitive web interface hides much of technicalities from the user, allowing them to focus on the essence of the docking. The server supports both DNA and RNA and can easily be used in conjunction with the 3D-DART web server described in chapter 4. A Grid-enabled version of the HADDOCK server has been developed within the European FP7 e-Infrastructure eNMR project (http://www.enmr.eu). This grid connects many servers around Europe together, compensating in part for the increasing demand for computational power by the HADDOCK server. Since its launch, the HADDOCK web server has processed over 5300 requests from users all around the world. In a similar fashion, the 3D-DART web server had dealt with over 2900 requests for custom-built DNA structural models, illustrating the demand for both services by the community.

Apart from the web server, HADDOCK has also been integrated into the CcpNmr Analysis software suite (http://www.ccpn.ac.uk, (334)) a popular NMR processing and analysis software. CCPN provides the underlying, unified, data model, which also serves as connection hub for a wide range of NMR related software packages. The ability of HADDOCK to exchange data with the CcpNmr Analysis software allows NMR spectroscopists to prepare a docking run based on for instance NMR structures and experimental data from within the familiar environment of CcpNmr Analysis and launch the docking on our HADDOCK server with single mouse click.

In conclusion, we are still far from a unified docking method that will give highly accurate results in an ab initio fashion and one may speculate whether this will ever be the case. Until that time, the use of experimental information will be essential to assemble the correct interfaces, deal with ever-larger conformational changes, design new experiments and/or validate the docking results. This clearly matches with the philosophy of HADDOCK as a practical docking method made to be an integral part of a research workflow. It has already demonstrated its usefulness for protein-protein docking and is now pioneering the field of protein-DNA docking. We certainly hope that the tools and methods described in this thesis will further contribute to the establishment of HADDOCK as a general biomolecular docking method.

# References

1. Adesokan, A.A., Roberts, V.A., Lee, K.W., Lins, R.D. and Briggs, J.M. (2003) Prediction of HIV-1 Integrase/Viral DNA Interactions in the Catalytic Domain by Fast Molecular Docking. J. Med. Chem., 47, 821-828.
2. Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M. and Harrison, S.C. (1988) Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. Science, 242, 899-907.
3. Ahmad, S., Gromiha, M.M. and Sarai, A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics (Oxford, England), 20, 477-486.
4. Ahmad, S. and Sarai, A. (2004) Moment-based prediction of DNA-binding proteins. J Mol Biol, 341, 65-71.
5. Ahmad, S. and Sarai, A. (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics, 6, 33.
6. Ahmad, S., Kono, H., Arauzo-Bravo, M.J. and Sarai, A. (2006) ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. Nucleic acids research, 34, W124-127.
7. Allemann, R.K. and Egli, M. (1997) DNA recognition and bending. Chem Biol, 4, 643-650.
8. Aloy, P., Moont, G., Gabb, H.A., Querol, E., Aviles, F.X. and Sternberg, M.J. (1998) Modelling repressor proteins docking to DNA. Proteins, 33, 535-549.
9. Aloy, P., Querol, E., Aviles, F.X. and Sternberg, M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol, 311, 395-408.
10. An, J., Nakama, T., Kubota, Y. and Sarai, A. (1998) 3DinSight: an integrated relational database and search tool for the structure, function and properties of biomolecules. Bioinformatics (Oxford, England), 14, 188-195.
11. Anderson, J.E., Ptashne, M. and Harrison, S.C. (1987) Structure of the repressor-operator complex of bacteriophage 434. Nature, 326, 846-852.
12. Anderson, W.F., Takeda, Y., Echols, H. and Matthews, B.W. (1979) The structure of a repressor: crystallographic data for the Cro regulatory protein of bacteriophage lambda. J Mol Biol, 130, 507-510.
13. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. N Biotechnol, 25, 195-203.
14. Argast, G.M., Stephens, K.M., Emond, M.J. and Monnat, R.J., Jr. (1998) I-PpoI and I-CreI homing site sequence degeneracy determined by random mutagenesis and sequential in vitro enrichment. J Mol Biol, 280, 345-353.
15. Ausiello, G., Cesareni, G. and Helmer-Citterich, M. (1997) ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure. Proteins, 28, 556-567.
16. Avery, O., MacLeod, C. and M., M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med., 79, 137-158.
17. Axelrod, J.D., Majors, J. and Brandriss, M.C. (1991) Proline-independent binding of PUT3 transcriptional activator protein detected by footprinting in vivo. Mol Cell Biol, 11, 564-567.

18. Bahar, I. and Rader, A.J. (2005) Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol, 15, 586-592.
19. Barkley, M.D. and Bourgeois, S. (1978) In Miller, J. H. and Reznikoff, W. S. (eds.), The operon. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp. 177-200.
20. Bastard, K., Thureau, A., Lavery, R. and Prevost, C. (2003) Docking macromolecules with flexible segments. J Comput Chem, 24, 1910-1920.
21. Bastia, D. (1996) Structural aspects of protein-DNA interactions as revealed by conversion of the interacting protein into a sequence-specific cross-linking agent or a chemical nuclease. Structure, 4, 661-664.
22. Becker, N.B. and Everaers, R. (2009) DNA nanomechanics: how proteins deform the double helix. J Chem Phys, 130, 135102.
23. Bedrosian, C.L. and Bastia, D. (1990) The DNA-binding domain of HPV-16 E2 protein interaction with the viral enhancer: protein-induced DNA bending and role of the nonconserved core sequence in binding site affinity. Virology, 174, 557-575.
24. Benos, P.V., Bulyk, M.L. and Stormo, G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? Nucleic acids research, 30, 4442-4451.
25. Benos, P.V., Lapedes, A.S. and Stormo, G.D. (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. Bioessays, 24, 466-475.
26. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2006) GenBank. Nucleic acids research, 34, D16-20.
27. Berg, O.G. and von Hippel, P.H. (1988) Selection of DNA binding sites by regulatory proteins. Trends Biochem Sci, 13, 207-211.
28. Berman, H., Henrick, K., Nakamura, H. and Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic acids research, 35, 301-303.
29. Berman, H.M. (1997) Crystal studies of B-DNA: the answers and the questions. Biopolymers, 44, 23-44.
30. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. Nucleic acids research, 28, 235-242.
31. Bessiere, D., Lacroix, C., Campagne, S., Ecochard, V., Guillet, V., Mourey, L., Lopez, F., Czaplicki, J., Demange, P., Milon, A. *et al*. (2008) Structure-function analysis of the THAP zinc finger of THAP1, a large C2CH DNA-binding module linked to Rb/E2F pathways. J Biol Chem, 283, 4352-4363.
32. Beveridge, D.L. and McConnell, K.J. (2000) Nucleic acids: theory and computer simulation, Y2K. Curr Opin Struct Biol, 10, 182-196.
33. Beveridge, D.L., Barreiro, G., Byun, K.S., Case, D.A., Cheatham, T.E., 3rd, Dixit, S.B., Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H. *et al*. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. Biophys J, 87, 3799-3813.
34. Bhattacharya, A., Tejero, R. and Montelione, G.T. (2007) Evaluating protein structures determined by structural genomics consortia. Proteins, 66, 778-795.
35. Biswas, S., Guharoy, M. and Chakrabarti, P. (2009) Dissection, residue conservation, and structural classification of protein-DNA interfaces. Proteins, 74, 643-654.
36. Boelens, R., Scheek, R.M., van Boom, J.H. and Kaptein, R. (1987) Complex of lac repressor headpiece with a 14 base-pair lac operator fragment studied by two-

dimensional nuclear magnetic resonance. J Mol Biol, 193, 213-216.
37. Bonvin, A.M., Vis, H., Breg, J.N., Burgering, M.J., Boelens, R. and Kaptein, R. (1994) Nuclear magnetic resonance solution structure of the Arc repressor using relaxation matrix calculations. J Mol Biol, 236, 328-341.
38. Bonvin, A.M., Boelens, R. and Kaptein, R. (2005) NMR analysis of protein interactions. Curr Opin Chem Biol, 9, 501-508.
39. Brandriss, M.C. (1987) Evidence for positive regulation of the proline utilization pathway in Saccharomyces cerevisiae. Genetics, 117, 429-435.
40. Breslauer, K.J., Frank, R., Blocker, H. and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. Proc Natl Acad Sci U S A, 83, 3746-3750.
41. Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S. *et al*. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta crystallographica, 54, 905-921.
42. Bushman, F.D., Anderson, J.E., Harrison, S.C. and Ptashne, M. (1985) Ethylation interference and X-ray crystallography identify similar interactions between 434 repressor and operator. Nature, 316, 651-653.
43. Caceres, R.A., Pauli, I., Timmers, L.F. and de Azevedo, W.F., Jr. (2008) Molecular recognition models: a challenge to overcome. Curr Drug Targets, 9, 1077-1083.
44. Cai, S., Zhu, L., Zhang, Z. and Chen, Y. (2007) Determination of the three-dimensional structure of the Mrf2-DNA complex using paramagnetic spin labeling. Biochemistry, 46, 4943-4950.
45. Calladine, C.R. and Drew, H.R. (1984) A base-centred explanation of the B-to-A transition in DNA. J Mol Biol, 178, 773-782.
46. Calladine, C.R. and Drew, H.R. (1986) Principles of sequence-dependent flexure of DNA. J Mol Biol, 192, 907-918.
47. Campbell, G., Deng, Y., Glimm, J., Wang, Y., Yu, Q., Eisenberg, M. and Grollman, A. (1996) Analysis and prediction of hydrogen bonding in protein-DNA complexes using parallel processors. J Chomp Chem, 17, 1712-1725.
48. Carter, P.J., Winter, G., Wilkinson, A.J. and Fersht, A.R. (1984) The use of double mutants to detect structural changes in the active site of the tyrosyl-tRNA synthetase (Bacillus stearothermophilus). Cell, 38, 835-840.
49. Chandler, D. (2005) Interfaces and the driving force of hydrophobic assembly. Nature, 437, 640-647.
50. Chandrasekaran, R.A. and Arnott, S. (1989) In Saenger, W. (ed.), Landolt-Börnstein Numerical Data and functional Relationships in Science and Technology. Springer-Verlag, Vol. VII/1b, pp. 31-170.
51. Charbonnier, S., Gallego, O. and Gavin, A.C. (2008) The social network of a cell: recent advances in interactome mapping. Biotechnol Annu Rev, 14, 1-28.
52. Chargaff, E. (1951) Some recent studies on the composition and structure of nucleic acids. J Cell Physiol Suppl, 38, 41-59.
53. Chargaff, E., Lipshitz, R., Green, C. and Hodes, M.E. (1951) The composition of the deoxyribonucleic acid of salmon sperm. J Biol Chem, 192, 223-230.
54. Chargaff, E., Lipshitz, R. and Green, C. (1952) Composition of the desoxypentose nucleic acids of four genera of sea-urchin. J Biol Chem, 195, 155-160.
55. Cheatham, T.E., 3rd. (2004) Simulation and modeling of nucleic acid structure, dynamics and interactions. Curr Opin Struct Biol, 14, 360-367.

56. Chen, R., Mintseris, J., Janin, J. and Weng, Z. (2003) A protein-protein docking benchmark. Proteins, 52, 88-91.
57. Chen, S., Gunasekera, A., Zhang, X., Kunkel, T.A., Ebright, R.H. and Berman, H.M. (2001) Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. J Mol Biol, 314, 75-82.
58. Chennubhotla, C., Rader, A.J., Yang, L.W. and Bahar, I. (2005) Elastic network models for understanding biomolecular machinery: from enzymes to supramolecular assemblies. Phys Biol, 2, S173-180.
59. Choo, Y. and Klug, A. (1997) Physical basis of a protein-DNA recognition code. Curr Opin Struct Biol, 7, 117-125.
60. Chuprina, V.P., Rullmann, J.A., Lamerichs, R.M., van Boom, J.H., Boelens, R. and Kaptein, R. (1993) Structure of the complex of lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restrained molecular dynamics. J Mol Biol, 234, 446-462.
61. Clark, A.T., Smith, K., Muhandiram, R., Edmondson, S.P. and Shriver, J.W. (2007) Carboxyl pK(a) values, ion pairs, hydrogen bonding, and the pH-dependence of folding the hyperthermophile proteins Sac7d and Sso7d. J Mol Biol, 372, 992-1008.
62. Crick, F.H., Barnett, L., Brenner, S. and Watts-Tobin, R.J. (1961) General nature of the genetic code for proteins. Nature, 192, 1227-1232.
63. Crick, F.H.C. (1958) Central Dogma of Molecular Biology. Nature, 227, 561-563.
64. Crothers, D.M. (1998) DNA curvature and deformation in protein-DNA complexes: a step in the right direction. Proc Natl Acad Sci U S A, 95, 15163-15165.
65. Curuksu, J., Zakrzewska, K. and Zacharias, M. (2008) Magnitude and direction of DNA bending induced by screw-axis orientation: influence of sequence, mismatches and abasic sites. Nucleic acids research, 36, 2268-2283.
66. Curuksu, J., Zacharias, M., Lavery, R. and Zakrzewska, K. (2009) Local and global effects of strong DNA bending induced during molecular dynamics simulations. Nucleic acids research, 37, 3766-3773.
67. Danielsen, M., Hinck, L. and Ringold, G.M. (1989) Two amino acids within the knuckle of the first zinc finger specify DNA response element activation by the glucocorticoid receptor. Cell, 57, 1131-1138.
68. Darwin, C. (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London.
69. Dauter, Z. (2006) Current state and prospects of macromolecular crystallography. Acta crystallographica, 62, 1-11.
70. Davis, I.W., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic acids research, 32, W615-619.
71. Day, W.H. and McMorris, F.R. (1992) Critical comparison of consensus methods for molecular sequences. Nucleic acids research, 20, 1093-1099.
72. de Groot, B.L., van Aalten, D.M., Scheek, R.M., Amadei, A., Vriend, G. and Berendsen, H.J. (1997) Prediction of protein conformational freedom from distance constraints. Proteins, 29, 240-251.
73. De Luca, L., Pedretti, A., Vistoli, G., Barreca, M.L., Villa, L., Monforte, P. and Chimirri, A. (2003) Analysis of the full-length integrase-DNA complex by a modified approach for DNA docking. Biochem Biophys Res Commun, 310, 1083-1088.

74. de Vries, S.J., van Dijk, A.D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T. and Bonvin, A.M. (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins, 69, 726-733.
75. DeLano, W.L. (2002) Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol, 12, 14-20.
76. Delarue, M. and Sanejouand, Y.H. (2002) Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. J Mol Biol, 320, 1011-1024.
77. Desjarlais, J.R. and Berg, J.M. (1992) Toward rules relating zinc finger protein sequences and DNA binding site preferences. Proc Natl Acad Sci U S A, 89, 7345-7349.
78. Devos, D. and Valencia, A. (2000) Practical limits of function prediction. Proteins, 41, 98-107.
79. Dickerson, R.E. (1989) Definitions and nomenclature of nucleic acid structure parameters. J Biomol Struct Dyn, 6, 627-634.
80. Dickerson, R.E., Goodsell, D. and Kopka, M.L. (1996) MPD and DNA bending in crystals and in solution. J Mol Biol, 256, 108-125.
81. Dickerson, R.E. and Chiu, T.K. (1997) Helix bending as a factor in protein/DNA recognition. Biopolymers, 44, 361-403.
82. Dickerson, R.E. (1998) DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic acids research, 26, 1906-1926.
83. Dixit, S.B. and Beveridge, D.L. (2005) Axis curvature and ligand induced bending in the CAP-DNA oligomers. Biophys. J., 88, L04-L06.
84. Dixit, S.B., Beveridge, D.L., Case, D.A., Cheatham, T.E., 3rd, Giudice, E., Lankas, F., Lavery, R., Maddocks, J.H., Osman, R., Sklenar, H. *et al*. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. Biophys J, 89, 3721-3740.
85. Dixit, S.B. and Beveridge, D.L. (2006) Structural bioinformatics of DNA: a web-based tool for the analysis of molecular dynamics results and structure prediction. Bioinformatics (Oxford, England), 22, 1007-1009.
86. Dominguez, C., Boelens, R. and Bonvin, A.M. (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. Journal of the American Chemical Society, 125, 1731-1737.
87. Dostal, L., Chen, C.Y., Wang, A.H. and Welfle, H. (2004) Partial B-to-A DNA transition upon minor groove binding of protein Sac7d monitored by Raman spectroscopy. Biochemistry, 43, 9600-9609.
88. Drew, H.R. and Travers, A.A. (1985) DNA bending and its relation to nucleosome positioning. J Mol Biol, 186, 773-790.
89. Dunbrack, R.L., Jr. (2002) Rotamer libraries in the 21st century. Curr Opin Struct Biol, 12, 431-440.
90. Dunn, R.K. and Kingston, R.E. (2007) Gene regulation in the postgenomic era: technology takes the wheel. Mol Cell, 28, 708-714.
91. Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol, 6, 197-208.
92. Eklund, J.L., Ulge, U.Y., Eastberg, J. and Monnat, R.J., Jr. (2007) Altered target site specificity variants of the I-PpoI His-Cys box homing endonuclease. Nucleic acids

research, 35, 5839-5850.
93. El Hassan, M.A. and Calladine, C.R. (1997) Conformational characteristics of DNA: emperical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. Phil. Trans. R. Soc. Lond. A., 355, 43-100.
94. El Hassan, M.A. and Calladine, C.R. (1998) Two distinct modes of protein-induced bending in DNA. J. Mol. Biol., 282, 331-343.
95. Ellison, E.L. and Vogt, V.M. (1993) Interaction of the intron-encoded mobility endonuclease I-PpoI with its target site. Mol Cell Biol, 13, 7531-7539.
96. Emekli, U., Schneidman-Duhovny, D., Wolfson, H.J., Nussinov, R. and Haliloglu, T. (2008) HingeProt: automated prediction of hinges in protein structures. Proteins, 70, 1219-1227.
97. Falcon, C.M. and Matthews, K.S. (2000) Operator DNA sequence variation enhances high affinity binding by hinge helix mutants of lactose repressor protein. Biochemistry, 39, 11074-11083.
98. Fan, L. and Roberts, V.A. (2006) Complex of linker histone H5 with the nucleosome and its implications for chromatin packing. Proc Natl Acad Sci U S A, 103, 8384-8389.
99. Fanelli, F. and Ferrari, S. (2006) Prediction of MEF2A-DNA interface by rigid body docking: a tool for fast estimation of protein mutational effects on DNA binding. J Struct Biol, 153, 278-283.
100. Farwer, J., Packer, M.J. and Hunter, C.A. (2006) Prediction of atomic structure from sequence for double helical DNA oligomers. Biopolymers, 81, 51-61.
101. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. *et al.* (2008) The Pfam protein families database. Nucleic acids research, 36, D281-288.
102. Flick, K.E., Jurica, M.S., Monnat, R.J., Jr. and Stoddard, B.L. (1998) DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. Nature, 394, 96-101.
103. Folmer, R.H., Nilges, M., Papavoine, C.H., Harmsen, B.J., Konings, R.N. and Hilbers, C.W. (1997) Refined structure, DNA binding studies, and dynamics of the bacteriophage Pf3 encoded single-stranded DNA binding protein. Biochemistry, 36, 9120-9135.
104. Frederick, C.A., Grable, J., Melia, M., Samudzi, C., Jen-Jacobson, L., Wang, B.C., Greene, P., Boyer, H.W. and Rosenberg, J.M. (1984) Kinked DNA in crystalline complex with EcoRI endonuclease. Nature, 309, 327-331.
105. Freemont, P.S., Friedman, J.M., Beese, L.S., Sanderson, M.R. and Steitz, T.A. (1988) Cocrystal structure of an editing complex of Klenow fragment with DNA. Proc Natl Acad Sci U S A, 85, 8924-8928.
106. Frishman, D., Mokrejs, M., Kosykh, D., Kastenmuller, G., Kolesov, G., Zubrzycki, I., Gruber, C., Geier, B., Kaps, A., Albermann, K. *et al.* (2003) The PEDANT genome database. Nucleic acids research, 31, 207-211.
107. Fuxreiter, M., Mezei, M., Simon, I. and Osman, R. (2005) Interfacial water as a "hydration fingerprint" in the noncognate complex of BamHI. Biophys J, 89, 903-911.
108. Galas, D.J. and Schmitz, A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding secificity. Nucleic acids research, 5, 3157-3170.
109. Galburt, E.A., Chadsey, M.S., Jurica, M.S., Chevalier, B.S., Erho, D., Tang, W.,

Monnat, R.J., Jr. and Stoddard, B.L. (2000) Conformational changes and cleavage by the homing endonuclease I-PpoI: a critical role for a leucine residue in the active site. J Mol Biol, 300, 877-887.

110. Gamsjaeger, R., Swanton, M.K., Kobus, F.J., Lehtomaki, E., Lowry, J.A., Kwan, A.H., Matthews, J.M. and Mackay, J.P. (2008) Structural and biophysical analysis of the DNA binding properties of myelin transcription factor 1. J Biol Chem, 283, 5158-5167.
111. Gane, P.J. and Dean, P.M. (2000) Recent advances in structure-based rational drug design. Curr Opin Struct Biol, 10, 401-404.
112. Gao, M. and Skolnick, J. (2009) From nonspecific DNA-protein encounter complexes to the prediction of DNA-protein interactions. PLoS Comput Biol, 5, e1000341.
113. Ginalski, K. (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol, 16, 172-177.
114. Glass, C.K. (1994) Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers. Endocr Rev, 15, 391-407.
115. Gorenstein, D.G., Schroeder, S.A., Fu, J.M., Metz, J.T., Roongta, V. and Jones, C.R. (1988) Assignments of 31P NMR resonances in oligodeoxyribonucleotides: origin of sequence-specific variations in the deoxyribose phosphate backbone conformation and the 31P chemical shifts of double-helical nucleic acids. Biochemistry, 27, 7223-7237.
116. Gorin, A.A., Zhurkin, V.B. and Olson, W.K. (1995) B-DNA twisting correlates with base-pair morphology. J Mol Biol, 247, 34-48.
117. Grzeskowiak, K., Goodsell, D.S., Kaczor-Grzeskowiak, M., Cascio, D. and Dickerson, R.E. (1993) Crystallographic analysis of C-C-A-A-G-C-T-T-G-G and its implications for bending in B-DNA. Biochemistry, 32, 8923-8931.
118. Gunther, S., Rother, K. and Frommel, C. (2006) Molecular flexibility in protein-DNA interactions. Biosystems, 85, 126-136.
119. Hall, B.A., Kaye, S.L., Pang, A., Perera, R. and Biggin, P.C. (2007) Characterization of protein conformational states by normal-mode frequencies. Journal of the American Chemical Society, 129, 11394-11401.
120. Halperin, I., Ma, B., Wolfson, H. and Nussinov, R. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins, 47, 409-443.
121. Hanson, C.L. and Robinson, C.V. (2004) Protein-nucleic acid interactions and the expanding role of mass spectrometry. J Biol Chem, 279, 24907-24910.
122. Hard, T. and Lundback, T. (1996) Thermodynamics of sequence-specific protein-DNA interactions. Biophys Chem, 62, 121-139.
123. Hard, T. (1999) NMR studies of protein-nucleic acid complexes: structures, solvation, dynamics and coupled protein folding. Q Rev Biophys, 32, 57-98.
124. Harison, S.C. (1991) A structural taxonomy of DNA-binding domains. Nature, 353, 715-719.
125. Harrison, S.C., Anderson, J.E., Koudelka, G.B., Mondragon, A., Subbiah, S., Wharton, R.P., Wolberger, C. and Ptashne, M. (1988) Recognition of DNA sequences by the repressor of bacteriophage 434. Biophys Chem, 29, 31-37.
126. Haussler, M.R., Whitfield, G.K., Haussler, C.A., Hsieh, J.C., Thompson, P.D., Selznick, S.H., Dominguez, C.E. and Jurutka, P.W. (1998) The nuclear vitamin D receptor: biological and molecular regulatory properties revealed. J Bone Miner Res, 13, 325-

349.
127. Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol, 288, 147-164.
128. Hershey, A. and Chase, M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol., 36, 39-56.
129. Hoffman, M.M., Khrapov, M.A., Cox, J.C., Yao, J., Tong, L. and Ellington, A.D. (2004) AANT: the Amino Acid-Nucleotide Interaction Database. Nucleic acids research, 32, D174-181.
130. Honig, B., Sharp, K. and Gilson, M. (1989) Electrostatic interactions in proteins. Prog Clin Biol Res, 289, 65-74.
131. Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996) Errors in protein structures. Nature, 381, 272.
132. Horton, N.C., Dorner, L.F. and Perona, J.J. (2002) Sequence selectivity and degeneracy of a restriction endonuclease mediated by DNA intercalation. Nat Struct Biol, 9, 42-47.
133. Hrmova, M. and Fincher, G.B. (2009) Functional genomics and structural biology in the definition of gene function. Methods Mol Biol, 513, 199-227.
134. Hubbard, S.J. and Thornton, J.M. (1993). Department of Biochemistry and Molecular Biology, University College London.
135. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S. and Sigrist, C.J. (2008) The 20 years of PROSITE. Nucleic acids research, 36, D245-249.
136. Hwang, S., Gou, Z. and Kuznetsov, I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics (Oxford, England), 23, 634-636.
137. Jamin, N. and Toma, F. (2001) NMR studies of protein-DNA interactions. Prog Nuc Mag Res Spec, 38, 83-114.
138. Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I. and Wodak, S.J. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins, 52, 2-9.
139. Janin, J. (2005) Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci, 14, 278-283.
140. Janin, J. (2007) The targets of CAPRI rounds 6-12. Proteins, 69, 699-703.
141. Jayaram, B. and Jain, T. (2004) The role of water in protein-DNA recognition. Annu Rev Biophys Biomol Struct, 33, 343-361.
142. Jones, S., Heyningen, P., Berman, H. and Thornton, J.M. (1999) Protein-DNA interactions: A structural analysis. J. Mol. Biol., 287, 877-896.
143. Jones, S., Barker, J.A., Nobeli, I. and Thornton, J.M. (2003) Using structural motif templates to identify proteins with DNA binding function. Nucleic acids research, 31, 2811-2823.
144. Jordan, S.R. and Pabo, C.O. (1988) Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions. Science, 242, 893-899.
145. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water. J. Chem. Phys., 79, 926-935.
146. Joseph-McCarthy, D. (1999) Computational approaches to structure-based ligand

design. Pharmacol Ther, 84, 179-191.
147. Kahsai, M.A., Martin, E., Edmondson, S.P. and Shriver, J.W. (2005) Stability and flexibility in the structure of the hyperthermophile DNA-binding protein Sac7d. Biochemistry, 44, 13500-13509.
148. Kalodimos, C.G., Biris, N., Bonvin, A.M., Levandoski, M.M., Guennuegues, M., Boelens, R. and Kaptein, R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. Science, 305, 386-389.
149. Kamphuis, M.B., Bonvin, A.M., Monti, M.C., Lemonnier, M., Munoz-Gomez, A., van den Heuvel, R.H., Diaz-Orejas, R. and Boelens, R. (2006) Model for RNA binding and the catalytic site of the RNase Kid of the bacterial parD toxin-antitoxin system. J Mol Biol, 357, 115-126.
150. Kaptein, R., Zuiderweg, E.R., Scheek, R.M., Boelens, R. and van Gunsteren, W.F. (1985) A protein structure from nuclear magnetic resonance data. lac repressor headpiece. J Mol Biol, 182, 179-182.
151. Kasinos, N., Lilley, G.A., Subbarao, N. and Haneef, I. (1992) A robust and efficient automated docking algorithm for molecular recognition. Protein Eng, 5, 69-75.
152. Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C. and Vakser, I.A. (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A, 89, 2195-2199.
153. Kelly, W. and Stumpf, M. (2008) Protein-protein interactions: from global to local analyses. Curr Opin Biotechnol, 19, 396-403.
154. Kisters-Woike, B., Lehming, N., Sartorius, J., von Wilcken-Bergmann, B. and Muller-Hill, B. (1991) A model of the lac repressor-operator complex based on physical and genetic data. Eur J Biochem, 198, 411-419.
155. Knegtel, R.M., Antoon, J., Rullmann, C., Boelens, R. and Kaptein, R. (1994) MONTY: a Monte Carlo approach to protein-DNA recognition. J Mol Biol, 235, 318-324.
156. Knegtel, R.M., Boelens, R. and Kaptein, R. (1994) Monte Carlo docking of protein-DNA complexes: incorporation of DNA flexibility and experimental data. Protein Eng, 7, 761-767.
157. Knegtel, R.M., Fogh, R.H., Ottleben, G., Ruterjans, H., Dumoulin, P., Schnarr, M., Boelens, R. and Kaptein, R. (1995) A model for the LexA repressor DNA complex. Proteins, 21, 226-236.
158. Kono, H. and Sarai, A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. Proteins, 35, 114-131.
159. Kopke Salinas, R., Folkers, G.E., Bonvin, A.M., Das, D., Boelens, R. and Kaptein, R. (2005) Altered specificity in DNA binding by the lac repressor: a mutant lac headpiece that mimics the gal repressor. Chembiochem, 6, 1628-1637.
160. Kosikov, K.M., Gorin, A.A., Zhurkin, V.B. and Olson, W.K. (1999) DNA stretching and compression: large-scale simulations of double helical structures. J Mol Biol, 289, 1301-1326.
161. Koszewski, N.J., Reinhardt, T.A. and Horst, R.L. (1996) Vitamin D receptor interactions with the murine osteopontin response element. J Steroid Biochem Mol Biol, 59, 377-388.
162. Koudelka, G.B., Harbury, P., Harrison, S.C. and Ptashne, M. (1988) DNA twisting and the affinity of bacteriophage 434 operator for bacteriophage 434 repressor. Proc Natl Acad Sci U S A, 85, 4633-4637.

163. Koudelka, G.B. and Carlson, P. (1992) DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. Nature, 355, 89-91.
164. Koudelka, G.B. and Lam, C.Y. (1993) Differential recognition of OR1 and OR3 by bacteriophage 434 repressor and Cro. J Biol Chem, 268, 23812-23817.
165. Koudelka, G.B. (1998) Recognition of DNA structure by 434 repressor. Nucleic acids research, 26, 669-675.
166. Kumar, M., Gromiha, M.M. and Raghava, G.P. (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics, 8, 463.
167. Kuntz, I.D. (1992) Structure-based strategies for drug design and discovery. Science, 257, 1078-1082.
168. Kuznetsov, I.B., Gou, Z., Li, R. and Hwang, S. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins, 64, 19-27.
169. Lafontaine, I. and Lavery, R. (1999) Collective variable modelling of nucleic acids. Curr Opin Struct Biol, 9, 170-176.
170. Lafontaine, I. and Lavery, R. (2000) Optimization of nucleic acid sequences. Biophys J, 79, 680-685.
171. Lafontaine, I. and Lavery, R. (2001) ADAPT: A molecular mechanics approach for studying the structural properties of long DNA sequences. Biopolymers, 56, 292-310.
172. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860-921.
173. Lankas, F., Sponer, J., Langowski, J. and Cheatham, T.E., 3rd. (2004) DNA deformability at the base pair level. Journal of the American Chemical Society, 126, 4124-4125.
174. Lanman, J. and Prevelige, P.E., Jr. (2004) High-sensitivity mass spectrometry for imaging subunit interactions: hydrogen/deuterium exchange. Curr Opin Struct Biol, 14, 181-188.
175. Larson, C.L. and Verdine, G.L. (1996) In S.M., H. (ed.), Nucleic Acids. Oxfort University Press, New York, pp. 324-346.
176. Laskowski, R.A., Moss, D. and Thornton, J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst., 283-291.
177. Lavery, R. and Sklenar, H. (1988) The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. J Biomol Struct Dyn, 6, 63-91.
178. Lavery, R., Zakrzewska, K. and Sklenar, H. (1995) JUMNA (junction minimisation of nucleic acids). Comp. Phys. Comm., 91, 135-158.
179. Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. Nucleic acids research.
180. Lawson, C.L., Swigon, D., Murakami, K.S., Darst, S.A., Berman, H.M. and Ebright, R.H. (2004) Catabolite activator protein: DNA binding and transcription activation. Curr Opin Struct Biol, 14, 10-20.
181. Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A. and Wright, P.E. (1989) Three-dimensional solution structure of a single zinc finger DNA-binding domain. Science, 245, 635-637.
182. Lee, M.S., Kliewer, S.A., Provencal, J., Wright, P.E. and Evans, R.M. (1993) Structure of

the retinoid X receptor alpha DNA binding domain: a helix required for homodimeric DNA binding. Science, 260, 1117-1121.
183. Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A. and Brasseur, R. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins, 61, 258-271.
184. Lengauer, T. and Rarey, M. (1996) Computational methods for biomolecular docking. Curr Opin Struct Biol, 6, 402-406.
185. Letunic, I., Doerks, T. and Bork, P. (2009) SMART 6: recent updates and new developments. Nucleic acids research, 37, D229-232.
186. Levene, P. (1919) The structure of yeast nucleic acid. J. Biol. Chem., 40, 415.
187. Lindahl, E. and Delarue, M. (2005) Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization. Nucleic acids research, 33, 4496-4506.
188. Linge, J.P., Williams, M.A., Spronk, C.A., Bonvin, A.M. and Nilges, M. (2003) Refinement of protein structures in explicit solvent. Proteins, 50, 496-506.
189. Liu, W., Vierke, G., Wenke, A.K., Thomm, M. and Ladenstein, R. (2007) Crystal structure of the archaeal heat shock regulator from Pyrococcus furiosus: a molecular chimera representing eukaryal and bacterial features. J Mol Biol, 369, 474-488.
190. Liu, Z., Mao, F., Guo, J.T., Yan, B., Wang, P., Qu, Y. and Xu, Y. (2005) Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. Nucleic acids research, 33, 546-558.
191. Liu, Z., Guo, J.T., Li, T. and Xu, Y. (2008) Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. Proteins, 72, 1114-1124.
192. Lorenz, M.G. and Wackernagel, W. (1994) Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev, 58, 563-602.
193. Lou, Y.C., Wei, S.Y., Rajasekaran, M., Chou, C.C., Hsu, H.M., Tai, J.H. and Chen, C. (2009) NMR structural analysis of DNA recognition by a novel Myb1 DNA-binding domain in the protozoan parasite Trichomonas vaginalis. Nucleic acids research, 37, 2381-2394.
194. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic acids research, 31, 5108-5121.
195. Lu, X.J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc, 3, 1213-1227.
196. Luque, I. and Freire, E. (2000) Structural stability of binding sites: consequences for binding affinity and allosteric effects. Proteins, Suppl 4, 63-71.
197. Luscombe, N.M., Austin, S.E., Berman, H.M. and Thornton, J.M. (2000) An overview of the structures of protein-DNA complexes. Genome biology, 1, online.
198. Luscombe, N.M., Laskowski, R.A. and Thornton, J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic acids research, 29, 2860-2874.
199. Lustig, B. and Jernigan, R.L. (1995) Consistencies of individual DNA base-amino acid interactions in structures and sequences. Nucleic acids research, 23, 4707-4711.
200. Lutter, L.C., Halvorson, H.R. and Calladine, C.R. (1996) Topological measurement of protein-induced DNA bend angles. J Mol Biol, 261, 620-633.
201. Macke, T. and Case, D.A. (1998) In Leontes, N. B. and SantaLucia, J. J. (eds.),

Molecular Modeling of Nucleic Acids. American Chemcial Society, Washington DC, pp. 379-393.
202. Mader, S., Kumar, V., de Verneuil, H. and Chambon, P. (1989) Three amino acids of the oestrogen receptor are essential to its ability to distinguish an oestrogen from a glucocorticoid-responsive element. Nature, 338, 271-274.
203. Majorek, K.A. and Bujnicki, J.M. (2009) Modeling of Escherichia coli Endonuclease V structure in complex with DNA. J Mol Model, 15, 173-182.
204. Mandel-Gutfreund, Y., Schueler, O. and Margalit, H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J Mol Biol, 253, 370-382.
205. Mandel-Gutfreund, Y. and Margalit, H. (1998) Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. Nucleic acids research, 26, 2306-2312.
206. Mandel-Gutfreund, Y., Margalit, H., Jernigan, R.L. and Zhurkin, V.B. (1998) A role for CH...O interactions in protein-DNA recognition. J Mol Biol, 277, 1129-1140.
207. Manke, T., Bringas, R. and Vingron, M. (2003) Correlating protein-DNA and protein-protein interaction networks. J Mol Biol, 333, 75-85.
208. Marczak, J.E. and Brandriss, M.C. (1989) Isolation of constitutive mutations affecting the proline utilization pathway in Saccharomyces cerevisiae and molecular analysis of the PUT3 transcriptional activator. Mol Cell Biol, 9, 4696-4705.
209. Marczak, J.E. and Brandriss, M.C. (1991) Analysis of constitutive and noninducible mutations of the PUT3 transcriptional activator. Mol Cell Biol, 11, 2609-2619.
210. Maris, A.E., Sawaya, M.R., Kaczor-Grzeskowiak, M., Jarvis, M.R., Bearson, S.M., Kopka, M.L., Schroder, I., Gunsalus, R.P. and Dickerson, R.E. (2002) Dimerization allows DNA target site recognition by the NarL response regulator. Nat Struct Biol, 9, 771-778.
211. Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S. and Miller, J.H. (1994) Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. J Mol Biol, 240, 421-433.
212. Matthew, J.B. and Ohlendorf, D.H. (1985) Electrostatic deformation of DNA by a DNA-binding protein. J Biol Chem, 260, 5860-5862.
213. Matthews, B.W. (1988) Protein-DNA interaction. No code for recognition. Nature, 335, 294-295.
214. May, A. and Zacharias, M. (2008) Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. Proteins, 70, 794-809.
215. Mazur, J., Sarai, A. and Jernigan, R.L. (1989) Sequence dependence of the B-A conformational transition of DNA. Biopolymers, 28, 1223-1233.
216. Melquiond, A.S.J. and Bonvin, A.M.J.J. (2009) In Zacharias, M. (ed.), Protein-protein complexes: analysis, modelling and drug design. Imperial College Press.
217. Mendel, J.G. (1866) Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, 4, 3-47.
218. Mendez, R., Leplae, R., De Maria, L. and Wodak, S.J. (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins, 52, 51-67.
219. Miller, J.C. and Pabo, C.O. (2001) Rearrangement of side-chains in a Zif268 mutant

highlights the complexities of zinc finger-DNA recognition. J Mol Biol, 313, 309-315.
220. Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. and Weng, Z. (2005) Protein-Protein Docking Benchmark 2.0: an update. Proteins, 60, 214-216.
221. Moitessier, N., Westhof, E. and Hanessian, S. (2006) Docking of aminoglycosides to hydrated and flexible RNA. J Med Chem, 49, 1023-1033.
222. Mondragon, A. and Harrison, S.C. (1991) The phage 434 Cro/OR1 complex at 2.5 A resolution. J Mol Biol, 219, 321-334.
223. Moreira, I.S., Fernandes, P.A. and Ramos, M.J. (2007) Hot spots--a review of the protein-protein interface determinant amino-acid residues. Proteins, 68, 803-812.
224. Morrison, K.L. and Weiss, G.A. (2001) Combinatorial alanine-scanning. Curr Opin Chem Biol, 5, 302-307.
225. Munteanu, M.G., Vlahovicek, K., Parthasaraty, S., Simon, I. and Pongor, S. (1998) Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena. Trends Biochem. Sci., 23, 341-346.
226. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol, 247, 536-540.
227. Muscarella, D.E., Ellison, E.L., Ruoff, B.M. and Vogt, V.M. (1990) Characterization of I-Ppo, an intron-encoded endonuclease that mediates homing of a group I intron in the ribosomal DNA of Physarum polycephalum. Mol Cell Biol, 10, 3386-3396.
228. Nadassy, K., Wodak, S.J. and Janin, J. (1999) Structural features of protein-nucleic acid recognition sites. Biochemistry, 38, 1999-2017.
229. Nardelli, J., Gibson, T.J., Vesque, C. and Charnay, P. (1991) Base sequence discrimination by zinc-finger DNA-binding domains. Nature, 349, 175-178.
230. Nekludova, L. and Pabo, C.O. (1994) Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. Proc Natl Acad Sci U S A, 91, 6948-6952.
231. Nelson, C.C., Hendy, S.C., Faris, J.S. and Romaniuk, P.J. (1996) Retinoid X receptor alters the determination of DNA binding specificity by the P-box amino acids of the thyroid hormone receptor. J Biol Chem, 271, 19464-19474.
232. Nilges, M. and O'Donogheu, S. (1998) Ambiguous NOEs and automated NOE assignment. Proc Nucl Magn Reson Spectrosc, 32, 107-139.
233. Noy, A., Perez, A., Lankas, F., Javier Luque, F. and Orozco, M. (2004) Relative flexibility of DNA and RNA: a molecular dynamics study. J Mol Biol, 343, 627-638.
234. Ofran, Y., Mysore, V. and Rost, B. (2007) Prediction of DNA-binding residues from sequence. Bioinformatics (Oxford, England), 23, i347-353.
235. Okroj, M., Jenkins, H.T., Herbert, A.P., Barlow, P.N. and Blom, A.M. (2008) Structural basis and functional effects of the interaction between complement inhibitor C4b-binding protein and DNA. Mol Immunol, 46, 62-69.
236. Olson, W.K. (1996) Simulating DNA at low resolution. Curr Opin Struct Biol, 6, 242-256.
237. Olson, W.K., Gorin, A.A., Lu, X.J., Hock, L.M. and Zhurkin, V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A, 95, 11163-11168.
238. Olson, W.K. and Zhurkin, V.B. (2000) Modeling DNA deformations. Curr Opin Struct Biol, 10, 286-297.
239. Ornstein, R.L. and Fresco, J.R. (1983) Correlation of Tm and sequence of DNA

duplexes with delta H computed by an improved empirical potential method. Biopolymers, 22, 1979-2000.
240. Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F. and Sigler, P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. Nature, 335, 321-329.
241. Pabo, C.O. and Lewis, M. (1982) The operator-binding domain of lambda repressor: structure and DNA recognition. Nature, 298, 443-447.
242. Pabo, C.O. and Sauer, R.T. (1984) Protein-DNA recognition. Annu Rev Biochem, 53, 293-321.
243. Pabo, C.O. and Sauer, R.T. (1992) Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem, 61, 1053-1095.
244. Pabo, C.O. and Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol, 301, 597-624.
245. Padmanabhan, S., Jimenez, M.A., Gonzalez, C., Sanz, J.M., Gimenez-Gallego, G. and Rico, M. (1997) Three-dimensional solution structure and stability of phage 434 Cro protein. Biochemistry, 36, 6424-6436.
246. Paillard, G. and Lavery, R. (2004) Analyzing protein-DNA recognition mechanisms. Structure, 12, 113-122.
247. Pan, C.Q., Landgraf, R. and Sigman, D.S. (1994) DNA-binding proteins as site-specific nucleases. Mol Microbiol, 12, 335-342.
248. Parraga, G., Young, L. and Klevit, R.E. (1989) Zinc-finger motifs and DNA binding. Trends Biochem Sci, 14, 398.
249. Patikoglou, G. and Burley, S.K. (1997) Eukaryotic transcription factor-DNA complexes. Annu Rev Biophys Biomol Struct, 26, 289-325.
250. Perez-Martin, J., Rojo, F. and de Lorenzo, V. (1994) Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. Microbiol Rev, 58, 268-290.
251. Peters, W.B., Edmondson, S.P. and Shriver, J.W. (2005) Effect of mutation of the Sac7d intercalating residues on the temperature dependence of DNA distortion and binding thermodynamics. Biochemistry, 44, 4794-4804.
252. Poulain, P., Saladin, A., Hartmann, B. and Prevost, C. (2008) Insights on protein-DNA recognition by coarse grain modelling. J Comput Chem, 29, 2582-2592.
253. Prabakaran, P., Siebers, J.G., Ahmad, S., Gromiha, M.M., Singarayan, M.G. and Sarai, A. (2006) Classification of protein-DNA complexes based on structural descriptors. Structure, 14, 1355-1367.
254. Prevost, C., Takahashi, M. and Lavery, R. (2009) Deforming DNA: from physics to biology. Chemphyschem, 10, 1399-1404.
255. Qian, Y.Q., Billeter, M., Otting, G., Muller, M., Gehring, W.J. and Wuthrich, K. (1989) The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. Cell, 59, 573-580.
256. Rastinejad, F., Perlmann, T., Evans, R.M. and Sigler, P.B. (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. Nature, 375, 203-211.
257. Raumann, B.E., Rould, M.A., Pabo, C.O. and Sauer, R.T. (1994) DNA recognition by beta-sheets in the Arc repressor-operator crystal structure. Nature, 367, 754-757.
258. Reddy, C.K., Das, A. and Jayaram, B. (2001) Do water molecules mediate protein-

DNA recognition? J Mol Biol, 314, 619-632.
259. Repanas, K., Fuentes, G., Cohen, S.X., Bonvin, A.M. and Perrakis, A. (2009) Insights into the DNA cleavage mechanism of human LINE-1 retrotransposon endonuclease. Proteins, 74, 917-928.
260. Rhodes, D., Schwabe, J.W., Chapman, L. and Fairall, L. (1996) Towards an understanding of protein-DNA recognition. Philosophical transactions of the Royal Society of London, 351, 501-509.
261. Ritchie, D.W. (2008) Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci, 9, 1-15.
262. Roberts, V.A., Case, D.A. and Tsui, V. (2004) Predicting interactions of winged-helix transcription factors with DNA. Proteins, 57, 172-187.
263. Rohs, R., Sklenar, H. and Shakked, Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. Structure, 13, 1499-1509.
264. Romanuka, J., Folkers, G.E., Biris, N., Tishchenko, E., Wienk, H., Bonvin, A.M., Kaptein, R. and Boelens, R. (2009) Specificity and affinity of Lac repressor for the auxiliary operators O2 and O3 are explained by the structures of their protein-DNA complexes. J Mol Biol, 390, 478-489.
265. Rueda, M., Chacon, P. and Orozco, M. (2007) Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. Structure, 15, 565-575.
266. Rumpel, S., Razeto, A., Pillar, C.M., Vijayan, V., Taylor, A., Giller, K., Gilmore, M.S., Becker, S. and Zweckstetter, M. (2004) Structure and DNA-binding properties of the cytolysin regulator CylR2 from Enterococcus faecalis. EMBO J, 23, 3632-3642.
267. Russell, R.B., Sasieni, P.D. and Sternberg, M.J. (1998) Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol, 282, 903-918.
268. Sanchez, I.E., Dellarole, M., Gaston, K. and de Prat Gay, G. (2008) Comprehensive comparison of the interaction of the E2 master regulator with its cognate target DNA sites in 73 human papillomavirus types by sequence statistics. Nucleic acids research, 36, 756-769.
269. Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins, 9, 56-68.
270. Sandmann, C., Cordes, F. and Saenger, W. (1996) Structure model of a complex between the factor for inversion stimulation (FIS) and DNA: modeling protein-DNA complexes with dyad symmetry and known protein structures. Proteins, 25, 486-500.
271. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature, 265, 687-695.
272. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 74, 5463-5467.
273. Sarai, A. and Kono, H. (2005) Protein-DNA recognition patterns and predictions. Annu Rev Biophys Biomol Struct, 34, 379-398.
274. Sathyapriya, R. and Vishveshwara, S. (2004) Interaction of DNA with clusters of amino acids in proteins. Nucleic acids research, 32, 4109-4118.
275. Sathyapriya, R., Vijayabaskar, M.S. and Vishveshwara, S. (2008) Insights into protein-DNA interactions through structure network analysis. PLoS Comput Biol, 4,

e1000170.
276. Schildbach, J.F., Karzai, A.W., Raumann, B.E. and Sauer, R.T. (1999) Origins of DNA-binding specificity: role of protein contacts with the DNA backbone. Proc Natl Acad Sci U S A, 96, 811-817.
277. Schneider, B., Cohen, D.M., Schleifer, L., Srinivasan, A.R., Olson, W.K. and Berman, H.M. (1993) A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. Biophys J, 65, 2291-2303.
278. Schneider, B. and Berman, H.M. (1995) Hydration of the DNA bases is local. Biophys J, 69, 2661-2669.
279. Schneider, B., Neidle, S. and Berman, H.M. (1997) Conformations of the sugar-phosphate backbone in helical DNA crystal structures. Biopolymers, 42, 113-124.
280. Schneider, B., Patel, K. and Berman, H.M. (1998) Hydration of the phosphate group in double-helical DNA. Biophys J, 75, 2422-2434.
281. Schneidman-Duhovny, D., Nussinov, R. and Wolfson, H.J. (2004) Predicting molecular interactions in silico: II. Protein-protein and protein-drug docking. Curr Med Chem, 11, 91-107.
282. Schroeder, S.A., Fu, J.M., Jones, C.R. and Gorenstein, D.G. (1987) Assignment of phosphorus-31 and nonexchangeable proton resonances in a symmetrical 14 base pair lac pseudooperator DNA fragment. Biochemistry, 26, 3812-3821.
283. Schroeder, S.A., Roongta, V., Fu, J.M., Jones, C.R. and Gorenstein, D.G. (1989) Sequence-dependent variations in the 31P NMR spectra and backbone torsional angles of wild-type and mutant Lac operator fragments. Biochemistry, 28, 8292-8303.
284. Schultz, S.C., Shields, G.C. and Steitz, T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. Science, 253, 1001-1007.
285. Schwabe, J.W. (1997) The role of water in protein-DNA interactions. Curr Opin Struct Biol, 7, 126-134.
286. Seeman, N.C., Rosenberg, J.M. and Rich, A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A, 73, 804-808.
287. Segal, D.J. and Barbas, C.F., 3rd. (2000) Design of novel sequence-specific DNA-binding proteins. Curr Opin Chem Biol, 4, 34-39.
288. Sen, T.Z., Kloczkowski, A. and Jernigan, R.L. (2006) A DNA-centric look at protein-DNA complexes. Structure, 14, 1341-1342.
289. Shakked, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. and Sigler, P.B. (1994) Determinants of repressor/operator recognition from the structure of the trp operator binding site. Nature, 368, 469-473.
290. Shatsky, M., Nussinov, R. and Wolfson, H.J. (2002) Flexible protein alignment and hinge detection. Proteins, 48, 242-256.
291. Shatsky, M., Nussinov, R. and Wolfson, H.J. (2004) FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. J Comput Biol, 11, 83-106.
292. Siddiqui, A.H. and Brandriss, M.C. (1989) The Saccharomyces cerevisiae PUT3 activator protein associates with proline-specific upstream activation sequences. Mol Cell Biol, 9, 4706-4712.
293. Sierk, M.L. and Kleywegt, G.J. (2004) Deja vu all over again: finding and analyzing protein structure similarities. Structure, 12, 2103-2111.
294. Singh, S., Hager, M.H., Zhang, C., Griffith, B.R., Lee, M.S., Hallenga, K., Markley,

J.L. and Thorson, J.S. (2006) Structural insight into the self-sacrifice mechanism of enediyne resistance. ACS Chem Biol, 1, 451-460.
295. Slijper, M., Bonvin, A.M., Boelens, R. and Kaptein, R. (1996) Refined structure of lac repressor headpiece (1-56) determined by relaxation matrix calculations from 2D and 3D NOE data: change of tertiary structure upon binding to the lac operator. J Mol Biol, 259, 761-773.
296. Sotriffer, C.A., Flader, W., Winger, R.H., Rode, B.M., Liedl, K.R. and Varga, J.M. (2000) Automated docking of ligands to antibodies: methods and applications. Methods, 20, 280-291.
297. Spronk, C.A., Bonvin, A.M., Radha, P.K., Melacini, G., Boelens, R. and Kaptein, R. (1999) The solution structure of Lac repressor headpiece 62 complexed to a symmetrical lac operator. Structure, 7, 1483-1492.
298. Stawiski, E.W., Gregoret, L.M. and Mandel-Gutfreund, Y. (2003) Annotating nucleic acid-binding function based on protein structure. J Mol Biol, 326, 1065-1079.
299. Steitz, T.A., Richmond, T.J., Wise, D. and Engelman, D. (1974) The lac repressor protein: molecular shape, subunit structure, and proposed model for operator interaction based on structural studies of microcrystals. Proc Natl Acad Sci U S A, 71, 593-597.
300. Stormo, G.D. (2000) DNA binding sites: representation and discovery. Bioinformatics (Oxford, England), 16, 16-23.
301. Suck, D., Lahm, A. and Oefner, C. (1988) Structure refined to 2A of a nicked DNA octanucleotide complex with DNase I. Nature, 332, 464-468.
302. Suhre, K. and Sanejouand, Y.H. (2004) On the potential of normal-mode analysis for solving difficult molecular-replacement problems. Acta crystallographica, 60, 796-799.
303. Suzuki, M. and Yagi, N. (1994) DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. Proc Natl Acad Sci U S A, 91, 12357-12361.
304. Suzuki, M., Brenner, S.E., Gerstein, M. and Yagi, N. (1995) DNA recognition code of transcription factors. Protein Eng, 8, 319-328.
305. Suzuki, M. and Yagi, N. (1995) Stereochemical basis of DNA bending by transcription factors. Nucleic acids research, 23, 2083-2091.
306. Suzuki, M., Loakes, D. and Yagi, N. (1996) DNA conformation and its changes upon binding transcription factors. Adv Biophys, 32, 53-72.
307. Suzuki, M., Yagi, N. and Finch, J.T. (1996) Role of base-backbone and base-base interactions in alternating DNA conformations. FEBS Lett, 379, 148-152.
308. Suzuki, M., Amano, N., Kakinuma, J. and Tateno, M. (1997) Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. J Mol Biol, 274, 421-435.
309. Tama, F. and Sanejouand, Y.H. (2001) Conformational change of proteins arising from normal mode calculations. Protein Eng, 14, 1-6.
310. Thorn, K.S. and Bogan, A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics (Oxford, England), 17, 284-285.
311. Tjong, H. and Zhou, H.X. (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic acids research, 35, 1465-1477.
312. Tolstorukov, M.Y., Jernigan, R.L. and Zhurkin, V.B. (2004) Protein-DNA hydrophobic

recognition in the minor groove is facilitated by sugar switching. J Mol Biol, 337, 65-76.
313. Travers, A.A. (1989) DNA conformation and protein binding. Annu Rev Biochem, 58, 427-452.
314. Travers, A.A. (1992) The reprogramming of transcriptional competence. Cell, 69, 573-575.
315. Travers, A.A. (2004) The structural basis of DNA flexibility. Philos Transact A Math Phys Eng Sci, 362, 1423-1438.
316. Tung, C.S. and Carter, E.S., 2nd. (1994) Nucleic acid modeling tool (NAMOT): an interactive graphic tool for modeling nucleic acid structures. Comput Appl Biosci, 10, 427-433.
317. Tzou, W.S. and Hwang, M.J. (1999) Modeling helix-turn-helix protein-induced DNA bending with knowledge-based distance restraints. Biophys J, 77, 1191-1205.
318. Umesono, K. and Evans, R.M. (1989) Determinants of target gene specificity for steroid/thyroid hormone receptors. Cell, 57, 1139-1146.
319. Uversky, V.N., Gillespie, J.R. and Fink, A.L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins, 41, 415-427.
320. Vajda, S., Sippl, M. and Novotny, J. (1997) Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol, 7, 222-228.
321. Vajda, S. (2005) Classification of protein complexes based on docking difficulty. Proteins, 60, 176-180.
322. van Aalten, D.M., Conn, D.A., de Groot, B.L., Berendsen, H.J., Findlay, J.B. and Amadei, A. (1997) Protein dynamics derived from clusters of crystal structures. Biophys J, 73, 2891-2896.
323. van Dijk, A.D., Boelens, R. and Bonvin, A.M. (2005) Data-driven docking for the study of biomolecular complexes. The FEBS journal, 272, 293-312.
324. van Dijk, A.D., de Vries, S.J., Dominguez, C., Chen, H., Zhou, H.X. and Bonvin, A.M. (2005) Data-driven docking: HADDOCK's adventures in CAPRI. Proteins, 60, 232-238.
325. van Dijk, A.D. and Bonvin, A.M. (2006) Solvated docking: introducing water into the modelling of biomolecular complexes. Bioinformatics (Oxford, England), 22, 2340-2347.
326. van Dijk, M., van Dijk, A.D., Hsu, V., Boelens, R. and Bonvin, A.M. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic acids research, 34, 3317-3325.
327. van Dijk, M. and Bonvin, A.M. (2008) A protein-DNA docking benchmark. Nucleic acids research, 36, e88.
328. van Dijk, M. and Bonvin, A.M. (2009) 3D-DART: a DNA structure modelling server. Nucleic acids research.
329. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. Science, 291, 1304-1351.
330. Vershon, A.K., Kelley, R.D. and Sauer, R.T. (1989) Sequence-specific binding of arc repressor to DNA. Effects of operator mutations and modifications. J Biol Chem, 264, 3267-3273.
331. Vieth, M., Hirst, J.D., Kolinski, A. and Brooks, C.L.I. (1998) Assessing energy functions for flexible docking. J Chomp Chem, 19, 1612-1622.

332. Vlahovicek, K., Kajan, L. and Pongor, S. (2003) DNA analysis servers: plot.it, bend.it, model.it and IS. Nucleic acids research, 31, 3686-3687.
333. Volpon, L., D'Orso, I., Young, C.R., Frasch, A.C. and Gehring, K. (2005) NMR structural study of TcUBP1, a single RRM domain protein from Trypanosoma cruzi: contribution of a beta hairpin to RNA binding. Biochemistry, 44, 3708-3717.
334. Vranken, W.F., Boucher, W., Stevens, T.J., Fogh, R.H., Pajon, A., Llinas, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E.D. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins, 59, 687-696.
335. Vries de, H. (1905) Species and variation, their origin by mutation; lectures delivered at the University of California by Hugo De Vries. . The Open court publishing company, Chicago.
336. Wallace, A.C., Laskowski, R.A. and Thornton, J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng, 8, 127-134.
337. Walters, K.J., Dayie, K.T., Reece, R.J., Ptashne, M. and Wagner, G. (1997) Structure and mobility of the PUT3 dimer. Nat Struct Biol, 4, 744-750.
338. Wang, A.H., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G. and Rich, A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature, 282, 680-686.
339. Wang, L. and Brown, S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic acids research, 34, W243-248.
340. Warwicker, J., Engelman, B.P. and Steitz, T.A. (1987) Electrostatic calculations and model-building suggest that DNA bound to CAP is sharply bent. Proteins, 2, 283-289.
341. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171, 737-738.
342. Weber, I.T., McKay, D.B. and Steitz, T.A. (1982) Two helix DNA binding motif of CAP found in lac repressor and gal repressor. Nucleic acids research, 10, 5085-5102.
343. Weber, I.T. and Steitz, T.A. (1984) Model of specific complex between catabolite gene activator protein and B-DNA suggested by electrostatic complementarity. Proc Natl Acad Sci U S A, 81, 3973-3977.
344. Werner, M.H., Gronenborn, A.M. and Clore, G.M. (1996) Intercalation, DNA kinking, and the control of transcription. Science, 271, 778-784.
345. Wharton, R.P., Brown, E.L. and Ptashne, M. (1984) Substituting an alpha-helix switches the sequence-specific DNA interactions of a repressor. Cell, 38, 361-369.
346. Wijmenga, S.S. and Buuren, B.N.M. (1998) The use of NMR methods for conformatonal studies of nucleic acids. Prog Nuc Mag Res Spec, 32, 287-387.
347. Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol, 297, 233-249.
348. Wittmayer, P.K., McKenzie, J.L. and Raines, R.T. (1998) Degenerate DNA recognition by I-PpoI endonuclease. Gene, 206, 11-21.
349. Woda, J., Schneider, B., Patel, K., Mistry, K. and Berman, H.M. (1998) An analysis of the relationship between hydration and protein-DNA interactions. Biophys J, 75, 2170-2177.
350. Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing

the protein structure-function paradigm. J Mol Biol, 293, 321-331.

351. Xu, B., Yang, Y., Liang, H. and Zhou, Y. (2009) An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. Proteins, 76, 718-730.
352. Yeh, C.S., Chen, F.M., Wang, J.Y., Cheng, T.L., Hwang, M.J. and Tzou, W.S. (2003) Directional shape complementarity at the protein-DNA interface. J Mol Recognit, 16, 213-222.
353. Young, M.A., Ravishanker, G., Beveridge, D.L. and Berman, H.M. (1995) Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. Biophys J, 68, 2454-2468.
354. Zacharias, M. and Sklenar, H. (1999) Harmonic modes as variables to approximately account for receptor flexibility in ligand-receptor docking simulations: application to DNA minor groove ligand complex. J Chomp Chem, 20, 287-399.
355. Zacharias, M. (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci, 12, 1271-1282.
356. Zheng, G., Lu, X.J. and Olson, W.K. (2009) Web 3DNA--a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. Nucleic acids research, 37, W240-246.
357. Zhou, H.X. and Qin, S. (2007) Interaction-site prediction for protein complexes: a critical assessment. Bioinformatics (Oxford, England), 23, 2203-2209.
358. Zhurkin, V.B., Lysov, Y.P. and Ivanov, V.I. (1979) Anisotropic flexibility of DNA and the nucleosomal structure. Nucleic acids research, 6, 1081-1096.
359. Zhurkin, V.B., Ulyanov, N.B., Gorin, A.A. and Jernigan, R.L. (1991) Static and statistical bending of DNA evaluated by Monte Carlo simulations. Proc Natl Acad Sci U S A, 88, 7046-7050.

# Appendix

## Appendix

**Table 1.** Average RMSD values from the target and fraction of native contacts for the docking starting from the unbound protein and a canonical B-DNA model. Statistics are given for the ten rigid-body refinement solutions that correspond to the ten best solutions of the top ranking cluster after semi-flexible refinement according to the HADDOCK score.

| | | RMSD (Å) | | | | | fnat[f] | CAPRI[g] |
|---|---|---|---|---|---|---|---|---|
| Complex | | Protein[a] | DNA[b] | Total[c] | Interface[d] | Ligand[e] | | -,*,** |
| **"easy"** | | | | | | | | |
| 1ea4 | 3 | $2.44_{0.20}$ | $3.41_{0.00}$ | $3.13_{0.15}$ | $3.23_{0.12}$ | $3.62_{1.14}$ | $0.29_{0.04}$ | 0,5,5 |
| 1ksy | 3 | $1.54_{0.52}$ | $2.17_{0.00}$ | $2.05_{0.26}$ | $2.47_{0.18}$ | $3.54_{1.21}$ | $0.30_{0.06}$ | 0,4,6 |
| 1rpe | 3 | $2.78_{0.34}$ | $2.86_{0.00}$ | $3.26_{0.18}$ | $2.91_{0.14}$ | $4.07_{0.79}$ | $0.22_{0.04}$ | 0,10,0 |
| 2c5r | 4 | $2.53_{0.22}$ | $0.85_{0.00}$ | $2.65_{0.26}$ | $2.24_{0.34}$ | $7.13_{1.75}$ | $0.47_{0.06}$ | 0,3,7 |
| 1tro | 3 | $3.51_{0.52}$ | $2.84_{0.00}$ | $3.68_{0.49}$ | $3.52_{0.48}$ | $6.14_{1.61}$ | $0.28_{0.05}$ | 0,10,0 |
| 1f4k | 3 | $3.07_{0.26}$ | $1.81_{0.00}$ | $3.14_{0.36}$ | $3.07_{0.36}$ | $4.33_{1.38}$ | $0.30_{0.06}$ | 0,5,5 |
| 1hjc | 2 | $1.77_{0.03}$ | $2.45_{0.00}$ | $2.76_{0.49}$ | $2.24_{0.24}$ | $7.86_{2.42}$ | $0.32_{0.04}$ | 0,8,2 |
| 1vrr | 3 | $3.19_{0.45}$ | $1.75_{0.00}$ | $3.48_{0.36}$ | $3.17_{0.13}$ | $8.64_{3.73}$ | $0.25_{0.07}$ | 0,9,1 |
| 3cro | 3 | $2.78_{0.25}$ | $2.71_{0.00}$ | $3.31_{0.21}$ | $3.04_{0.18}$ | $4.39_{0.89}$ | $0.19_{0.03}$ | 0,10,0 |
| 1r40 | 3 | $2.96_{0.38}$ | $2.41_{0.00}$ | $3.01_{0.30}$ | $3.21_{0.25}$ | $3.98_{0.86}$ | $0.20_{0.02}$ | 0,10,0 |
| 1mnn | 2 | $1.14_{0.00}$ | $1.66_{0.00}$ | $1.73_{0.24}$ | $2.04_{0.26}$ | $4.98_{3.68}$ | $0.23_{0.04}$ | 0,10,0 |
| 1wot | 3 | $4.66_{0.66}$ | $3.11_{0.00}$ | $5.00_{0.80}$ | $3.89_{0.52}$ | $4.83_{1.73}$ | $0.18_{0.05}$ | 1,9,0 |
| 1by4 | 3 | $4.18_{0.33}$ | $1.41_{0.00}$ | $3.70_{0.35}$ | $3.55_{0.23}$ | $3.87_{1.10}$ | $0.22_{0.03}$ | 0,10,0 |
| 1ddn | 5 | $3.84_{0.63}$ | $5.75_{0.00}$ | $4.75_{0.55}$ | $4.87_{0.51}$ | $7.29_{2.11}$ | $0.20_{0.03}$ | 2,8,0 |
| 1k79 | 2 | $1.48_{0.03}$ | $3.40_{0.00}$ | $3.34_{0.31}$ | $2.95_{0.24}$ | $5.52_{2.46}$ | $0.16_{0.03}$ | 0,10,0 |
| 1pt3 | 2 | $2.03_{0.10}$ | $1.89_{0.00}$ | $2.60_{0.17}$ | $2.94_{0.17}$ | $7.89_{0.98}$ | $0.14_{0.06}$ | 3,7,0 |
| 1cma | 3 | $3.17_{0.28}$ | $2.17_{0.00}$ | $3.20_{0.30}$ | $3.51_{0.18}$ | $5.65_{2.68}$ | $0.22_{0.02}$ | 0,10,0 |
| **"intermediate"** | | | | | | | | |
| 1fok | 2 | $1.72_{0.15}$ | $1.93_{0.00}$ | $2.24_{0.18}$ | $2.82_{0.24}$ | $7.43_{0.20}$ | $0.12_{0.00}$ | 0,10,0 |
| 1jj4 | 2 | $2.21_{0.07}$ | $3.19_{0.00}$ | $3.35_{0.15}$ | $3.73_{0.20}$ | $6.43_{1.08}$ | $0.15_{0.02}$ | 0,10,0 |
| 1h9t | 3 | $3.38_{0.33}$ | $3.32_{0.00}$ | $3.79_{0.33}$ | $3.72_{0.23}$ | $9.21_{4.29}$ | $0.11_{0.01}$ | 3,7,0 |
| 2irf | 2 | $4.67_{0.31}$ | $2.12_{0.00}$ | $4.51_{0.35}$ | $4.40_{0.53}$ | $7.04_{2.37}$ | $0.15_{0.06}$ | 4,6,0 |
| 1z63 | 2 | $2.67_{0.00}$ | $2.69_{0.00}$ | $3.58_{0.21}$ | $3.27_{0.02}$ | $12.14_{0.92}$ | $0.20_{0.03}$ | 0,10,0 |
| 7mht | 2 | $3.83_{0.07}$ | $2.51_{0.00}$ | $3.81_{0.07}$ | $6.70_{0.17}$ | $4.94_{1.17}$ | $0.08_{0.03}$ | 7,3,0 |
| 2fl3 | 2 | $4.10_{0.04}$ | $2.51_{0.00}$ | $4.32_{0.15}$ | $5.87_{0.18}$ | $8.96_{1.61}$ | $0.07_{0.02}$ | 9,1,0 |
| 1rva | 2 | $4.93_{0.20}$ | $4.09_{0.00}$ | $5.01_{0.23}$ | $5.45_{0.53}$ | $8.36_{3.69}$ | $0.05_{0.01}$ | 10,0,0 |
| 1eyu | 2 | $5.81_{0.43}$ | $3.93_{0.00}$ | $5.74_{0.47}$ | $5.67_{0.48}$ | $6.79_{2.47}$ | $0.05_{0.03}$ | 9,1,0 |
| 1b3t | 2 | $3.35_{0.01}$ | $3.56_{0.00}$ | $4.12_{0.16}$ | $4.87_{0.19}$ | $11.77_{3.42}$ | $0.09_{0.04}$ | 6,4,0 |
| 1diz | 2 | $0.99_{0.15}$ | $5.33_{0.00}$ | $3.38_{0.23}$ | $3.87_{0.27}$ | $19.46_{3.58}$ | $0.09_{0.03}$ | 7,3,0 |
| 1bdt | 3 | $5.33_{0.47}$ | $4.34_{0.00}$ | $5.27_{0.31}$ | $6.08_{0.34}$ | $6.29_{1.16}$ | $0.10_{0.02}$ | 7,3,0 |
| 1jto | 2 | $7.80_{1.37}$ | $3.49_{0.00}$ | $7.65_{1.35}$ | $6.09_{0.37}$ | $8.24_{2.04}$ | $0.08_{0.02}$ | 9,1,0 |
| 2fio | 2 | $1.46_{0.05}$ | $8.01_{0.00}$ | $6.80_{0.24}$ | $5.54_{0.07}$ | $8.68_{0.83}$ | $0.06_{0.02}$ | 10,0,0 |
| 1kc6 | 3 | $5.60_{0.19}$ | $4.46_{0.00}$ | $5.60_{0.15}$ | $4.33_{0.18}$ | $11.46_{2.14}$ | $0.12_{0.02}$ | 6,4,0 |

**Table 1.** Continued

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1zs4 | *2* | $6.02_{0.00}$ | $2.48_{0.00}$ | $6.10_{0.33}$ | $5.48_{0.28}$ | $10.14_{4.09}$ | $0.04_{0.02}$ | 10,0,0 |
| 1emh | *2* | $1.82_{0.07}$ | $4.38_{0.00}$ | $2.95_{0.18}$ | $3.75_{0.15}$ | $7.26_{2.04}$ | $0.05_{0.02}$ | 10,0,0 |
| **"difficult"** | | | | | | | | |
| 3bam | *3* | $6.23_{0.55}$ | $2.44_{0.00}$ | $6.17_{0.49}$ | $8.19_{0.42}$ | $8.76_{4.51}$ | $0.08_{0.04}$ | 10,0,0 |
| 1a74 | *2* | $1.75_{0.07}$ | $7.44_{0.00}$ | $5.11_{0.13}$ | $6.56_{0.14}$ | $8.60_{0.99}$ | $0.04_{0.01}$ | 10,0,0 |
| 4ktq | *2* | $2.45_{0.04}$ | $3.38_{0.00}$ | $3.25_{0.14}$ | $4.29_{0.17}$ | $8.62_{4.09}$ | $0.05_{0.01}$ | 10,0,0 |
| 1g9z | *2* | $4.68_{0.31}$ | $4.45_{0.00}$ | $5.69_{0.27}$ | $5.57_{0.25}$ | $8.67_{1.10}$ | $0.03_{0.01}$ | 10,0,0 |
| 1zme | *2* | $5.79_{0.03}$ | $4.28_{0.00}$ | $6.09_{0.26}$ | $4.92_{0.33}$ | $8.97_{1.85}$ | $0.09_{0.01}$ | 10,0,0 |
| 1qne | *2* | $1.74_{0.07}$ | $7.27_{0.00}$ | $4.49_{0.14}$ | $5.72_{0.12}$ | $6.50_{2.33}$ | $0.02_{0.00}$ | 10,0,0 |
| 1qrv | *3* | $7.18_{1.88}$ | $6.42_{0.00}$ | $8.06_{1.51}$ | $6.97_{1.08}$ | $11.34_{5.22}$ | $0.04_{0.01}$ | 10,0,0 |
| 1vas | *2* | $1.46_{0.06}$ | $5.87_{0.00}$ | $4.25_{0.17}$ | $5.00_{0.25}$ | $7.72_{3.45}$ | $0.01_{0.01}$ | 10,0,0 |
| 1azp | *2* | $3.61_{0.12}$ | $3.25_{0.00}$ | $4.79_{0.25}$ | $4.69_{0.24}$ | $9.53_{1.41}$ | $0.05_{0.03}$ | 10,0,0 |
| 1z9c | *3* | $8.734_{0.50}$ | $3.59_{0.00}$ | $7.75_{0.61}$ | $8.19_{0.53}$ | $12.64_{3.23}$ | $0.05_{0.02}$ | 10,0,0 |
| 103t | *2* | $2.893_{0.03}$ | $9.95_{0.00}$ | $7.63_{0.10}$ | $7.59_{0.06}$ | $12.01_{1.70}$ | $0.06_{0.01}$ | 10,0,0 |
| 2oaa | *2* | $8.342_{0.08}$ | $3.80_{0.00}$ | $8.17_{0.15}$ | $8.26_{0.14}$ | $19.82_{5.06}$ | $0.03_{0.01}$ | 10,0,0 |
| 1dfm | *3* | $7.672_{0.53}$ | $2.57_{0.00}$ | $8.39_{0.40}$ | $8.68_{0.31}$ | $17.12_{3.04}$ | $0.02_{0.01}$ | 10,0,0 |

Average RMSD values (Å) were calculated after superimposition of all heavy atoms belonging to the protein (a) the DNA (b), the full complex (c) and the interface (d). The ligand RMSD (e) was calculated by superimposition on all phosphate atoms of the DNA and subsequently on all Cα atoms of the proteins. (f) Fraction of native contact and (g) the number of no-, one- and two-star CAPRI solutions. Standard deviations are in subscript. The number of individual components docked is shown in italic after the PDB id.

**Table 2.** Average RMSD values from the target and fraction of native contacts for the docking starting from the unbound protein and a canonical B-DNA model. Statistics are given for the ten best solutions of the top ranking cluster after semi-flexible refinement according to the HADDOCK score.

| | | **RMSD (Å)** | | | | | **fnat[f]** | **CAPRI[g]** |
|---|---|---|---|---|---|---|---|---|
| **Complex** | | **Protein[a]** | **DNA[b]** | **Total[c]** | **Interface[d]** | **Ligand[e]** | | **-,*,**** |
| **"easy"** | | | | | | | | |
| 1ea4 | *3* | $2.35_{0.25}$ | $2.93_{0.18}$ | $2.70_{0.18}$ | $2.77_{0.18}$ | $2.46_{0.63}$ | $0.47_{0.05}$ | 0,0,10 |
| 1ksy | *3* | $1.77_{0.39}$ | $2.28_{0.16}$ | $2.20_{0.24}$ | $2.48_{0.19}$ | $3.46_{1.43}$ | $0.45_{0.04}$ | 0,2,8 |
| 1rpe | *3* | $2.87_{0.28}$ | $2.53_{0.21}$ | $2.90_{0.17}$ | $2.48_{0.15}$ | $3.25_{1.13}$ | $0.50_{0.06}$ | 0,1,9 |
| 2c5r | *4* | $2.63_{0.27}$ | $1.44_{0.23}$ | $2.78_{0.32}$ | $2.41_{0.36}$ | $7.17_{2.80}$ | $0.68_{0.04}$ | 0,5,5 |
| 1tro | *3* | $2.92_{0.28}$ | $2.73_{0.27}$ | $3.07_{0.28}$ | $2.90_{0.33}$ | $4.66_{1.51}$ | $0.38_{0.05}$ | 0,3,7 |
| 1f4k | *3* | $2.92_{0.11}$ | $2.24_{0.30}$ | $2.95_{0.25}$ | $2.86_{0.22}$ | $3.39_{1.24}$ | $0.36_{0.05}$ | 0,1,9 |
| 1hjc | *2* | $1.93_{0.18}$ | $2.44_{0.27}$ | $2.73_{0.55}$ | $2.11_{0.24}$ | $6.89_{2.90}$ | $0.51_{0.07}$ | 0,6,4 |
| 1vrr | *3* | $3.38_{0.11}$ | $2.03_{0.21}$ | $3.40_{0.08}$ | $2.82_{0.12}$ | $6.79_{1.54}$ | $0.40_{0.03}$ | 0,10,0 |
| 3cro | *3* | $2.47_{0.20}$ | $2.54_{0.15}$ | $2.87_{0.25}$ | $2.55_{0.24}$ | $3.32_{1.04}$ | $0.42_{0.05}$ | 0,1,9 |
| 1r40 | *3* | $3.02_{0.27}$ | $2.31_{0.21}$ | $2.92_{0.29}$ | $2.97_{0.24}$ | $3.45_{0.75}$ | $0.34_{0.04}$ | 0,2,8 |
| 1mnn | *2* | $1.67_{0.04}$ | $1.70_{0.11}$ | $1.93_{0.17}$ | $1.98_{0.23}$ | $4.90_{2.38}$ | $0.49_{0.04}$ | 0,1,9 |
| 1wot | *3* | $4.54_{0.62}$ | $3.46_{0.21}$ | $4.82_{0.63}$ | $3.75_{0.41}$ | $4.41_{1.34}$ | $0.34_{0.08}$ | 0,3,7 |

**Table 2.** Continued

| Complex | | RMSD (Å) Protein[a] | DNA[b] | Total[c] | Interface[d] | Ligand[e] | fnat[f] | CAPRI[g] -,*,** |
|---|---|---|---|---|---|---|---|---|
| 1by4 | 3 | $4.06_{0.38}$ | $1.81_{0.10}$ | $3.69_{0.39}$ | $3.51_{0.25}$ | $3.92_{0.94}$ | $0.36_{0.06}$ | 0,2,8 |
| 1ddn | 5 | $3.27_{0.74}$ | $5.08_{0.19}$ | $4.15_{0.57}$ | $4.18_{0.53}$ | $6.44_{1.62}$ | $0.42_{0.02}$ | 0,9,1 |
| 1k79 | 2 | $1.69_{0.06}$ | $3.21_{0.29}$ | $3.13_{0.39}$ | $2.76_{0.36}$ | $5.07_{2.06}$ | $0.37_{0.05}$ | 0,3,7 |
| 1pt3 | 2 | $2.25_{0.07}$ | $1.90_{0.16}$ | $2.69_{0.20}$ | $2.91_{0.22}$ | $7.37_{1.11}$ | $0.28_{0.07}$ | 0,7,3 |
| 1cma | 3 | $3.23_{0.21}$ | $1.89_{0.19}$ | $3.29_{0.27}$ | $3.44_{0.25}$ | $5.99_{2.43}$ | $0.40_{0.06}$ | 0,5,5 |
| **"intermediate"** | | | | | | | | |
| 1fok | 2 | $1.99_{0.06}$ | $1.86_{0.12}$ | $2.12_{0.08}$ | $2.34_{0.10}$ | $3.65_{0.75}$ | $0.26_{0.04}$ | 0,4,6 |
| 1jj4 | 2 | $4.76_{0.58}$ | $3.27_{0.28}$ | $4.45_{0.44}$ | $4.43_{0.24}$ | $5.51_{1.16}$ | $0.29_{0.04}$ | 0,8,2 |
| 1h9t | 3 | $2.62_{0.32}$ | $2.94_{0.16}$ | $3.07_{0.38}$ | $2.93_{0.33}$ | $7.41_{4.41}$ | $0.36_{0.07}$ | 0,9,1 |
| 2irf | 2 | $4.63_{0.30}$ | $2.26_{0.14}$ | $4.44_{0.45}$ | $4.26_{0.63}$ | $6.81_{2.57}$ | $0.29_{0.09}$ | 1,6,3 |
| 1z63 | 2 | $2.89_{0.03}$ | $2.91_{0.13}$ | $3.72_{0.20}$ | $3.14_{0.10}$ | $11.12_{0.66}$ | $0.39_{0.08}$ | 0,10,0 |
| 7mht | 2 | $3.69_{0.10}$ | $2.75_{0.20}$ | $3.68_{0.08}$ | $6.41_{0.19}$ | $4.76_{0.75}$ | $0.22_{0.05}$ | 0,10,0 |
| 2fl3 | 2 | $4.20_{0.05}$ | $2.25_{0.23}$ | $4.05_{0.14}$ | $5.35_{0.13}$ | $6.62_{1.30}$ | $0.24_{0.07}$ | 0,10,0 |
| 1rva | 2 | $3.96_{0.30}$ | $3.58_{0.23}$ | $4.02_{0.30}$ | $4.39_{0.38}$ | $6.35_{2.43}$ | $0.27_{0.02}$ | 1,8,1 |
| 1eyu | 2 | $4.23_{0.46}$ | $3.43_{0.19}$ | $4.29_{0.48}$ | $4.45_{0.48}$ | $5.01_{0.73}$ | $0.14_{0.05}$ | 0,8,2 |
| 1b3t | 2 | $3.37_{0.07}$ | $3.31_{0.23}$ | $3.77_{0.23}$ | $4.39_{0.30}$ | $9.80_{3.12}$ | $0.18_{0.03}$ | 5,5,0 |
| 1diz | 2 | $1.25_{0.12}$ | $5.37_{0.33}$ | $3.40_{0.36}$ | $3.30_{0.43}$ | $16.94_{3.76}$ | $0.26_{0.08}$ | 0,10,0 |
| 1bdt | 3 | $5.24_{0.53}$ | $4.04_{0.23}$ | $5.04_{0.42}$ | $5.77_{0.44}$ | $6.09_{1.40}$ | $0.25_{0.04}$ | 0,10,0 |
| 1jto | 2 | $7.81_{1.34}$ | $3.42_{0.08}$ | $7.59_{1.36}$ | $5.44_{0.39}$ | $6.86_{2.04}$ | $0.23_{0.04}$ | 1,9,0 |
| 2fio | 2 | $2.72_{0.25}$ | $8.22_{0.24}$ | $6.98_{0.27}$ | $5.37_{0.20}$ | $7.47_{1.13}$ | $0.14_{0.03}$ | 1,9,0 |
| 1kc6 | 3 | $4.96_{0.20}$ | $3.88_{0.18}$ | $4.98_{0.18}$ | $3.66_{0.14}$ | $10.45_{1.96}$ | $0.33_{0.02}$ | 0,10,0 |
| 1zs4 | 2 | $5.91_{0.30}$ | $2.49_{0.13}$ | $5.73_{0.31}$ | $4.71_{0.24}$ | $8.85_{4.17}$ | $0.15_{0.05}$ | 2,8,0 |
| 1emh | 2 | $1.92_{0.08}$ | $4.11_{0.22}$ | $2.97_{0.23}$ | $3.49_{0.25}$ | $7.89_{3.74}$ | $0.10_{0.05}$ | 3,7,0 |
| **"difficult"** | | | | | | | | |
| 3bam | 3 | $5.90_{0.42}$ | $2.69_{0.13}$ | $5.73_{0.43}$ | $7.50_{0.48}$ | $7.26_{3.07}$ | $0.19_{0.03}$ | 2,8,0 |
| 1a74 | 2 | $1.86_{0.13}$ | $6.79_{0.22}$ | $4.32_{0.20}$ | $5.51_{0.25}$ | $6.26_{1.19}$ | $0.24_{0.02}$ | 0,10,0 |
| 4ktq | 2 | $2.50_{0.09}$ | $3.36_{0.14}$ | $3.10_{0.16}$ | $3.72_{0.17}$ | $9.53_{2.45}$ | $0.23_{0.02}$ | 0,10,0 |
| 1g9z | 2 | $4.74_{0.17}$ | $4.19_{0.15}$ | $4.90_{0.20}$ | $4.32_{0.19}$ | $5.75_{0.63}$ | $0.21_{0.04}$ | 0,10,0 |
| 1zme | 2 | $5.90_{0.22}$ | $4.19_{0.39}$ | $5.82_{0.22}$ | $4.52_{0.32}$ | $7.67_{1.31}$ | $0.23_{0.05}$ | 1,9,0 |
| 1qne | 2 | $1.81_{0.12}$ | $5.93_{0.20}$ | $3.72_{0.20}$ | $4.06_{0.24}$ | $4.78_{2.21}$ | $0.14_{0.02}$ | 1,9,0 |
| 1qrv | 3 | $7.27_{1.67}$ | $5.80_{0.22}$ | $7.79_{1.48}$ | $6.40_{1.12}$ | $9.98_{5.26}$ | $0.12_{0.03}$ | 4,6,0 |
| 1vas | 2 | $1.71_{0.08}$ | $5.30_{0.31}$ | $3.73_{0.29}$ | $4.03_{0.44}$ | $6.18_{3.25}$ | $0.10_{0.03}$ | 4,6,0 |
| 1azp | 2 | $3.59_{0.16}$ | $2.79_{0.26}$ | $4.45_{0.48}$ | $4.08_{0.48}$ | $8.37_{1.96}$ | $0.12_{0.04}$ | 2,8,0 |
| 1z9c | 3 | $8.05_{0.45}$ | $3.38_{0.28}$ | $7.23_{0.50}$ | $7.54_{0.39}$ | $12.58_{2.84}$ | $0.13_{0.04}$ | 9,1,0 |
| 103t | 2 | $2.99_{0.15}$ | $9.77_{0.15}$ | $7.04_{0.19}$ | $6.91_{0.14}$ | $10.30_{2.20}$ | $0.12_{0.03}$ | 4,6,0 |
| 2oaa | 2 | $8.12_{0.16}$ | $3.42_{0.19}$ | $7.78_{0.16}$ | $7.61_{0.27}$ | $17.23_{3.35}$ | $0.17_{0.06}$ | 10,0,0 |
| 1dfm | 3 | $7.41_{0.55}$ | $2.58_{0.23}$ | $7.94_{0.45}$ | $8.26_{0.35}$ | $15.86_{3.09}$ | $0.04_{0.02}$ | 10,0,0 |

Average RMSD values (Å) were calculated after superimposition of all heavy atoms belonging to the protein (a) the DNA (b), the full complex (c) and the interface (d). The ligand RMSD (e) was calculated

by superimposition on all phosphate atoms of the DNA and subsequently on all Cα atoms of the proteins. (f) Fraction of native contact and (g) the number of no-, one- and two-star CAPRI solutions. Standard deviations are in subscript. The number of individual components docked is shown in italic after the PDB id.

**Table 3.** Average RMSD values from the target and fraction of native contacts for the docking starting from the unbound protein and an ensemble of custom-built DNA model. Statistics are given for the ten best solutions of the top ranking cluster after semi-flexible refinement according to the HADDOCK score.

| | | RMSD (Å) | | | | | fnat[f] | CAPRI[g] |
|---|---|---|---|---|---|---|---|---|
| **complex** | | **protein[a]** | **DNA[b]** | **Total[c]** | **Interface[d]** | **ligand[e]** | | **-,*,**** |
| **"easy"** | | | | | | | | |
| 1ea4 | *3* | $2.48_{0.20}$ | $2.76_{0.31}$ | $2.75_{0.21}$ | $2.63_{0.23}$ | $3.31_{0.86}$ | $0.53_{0.05}$ | 0,0,10 |
| 1ksy | *3* | $1.58_{0.15}$ | $2.03_{0.18}$ | $2.01_{0.19}$ | $2.14_{0.21}$ | $3.90_{2.02}$ | $0.54_{0.05}$ | 0,1,9 |
| 1rpe | *3* | $2.14_{0.20}$ | $1.70_{0.20}$ | $2.30_{0.27}$ | $2.00_{0.25}$ | $3.27_{1.26}$ | $0.57_{0.06}$ | 0,1,9 |
| 2c5r | *4* | $2.67_{0.36}$ | $1.45_{0.11}$ | $2.85_{0.31}$ | $2.45_{0.31}$ | $6.99_{3.40}$ | $0.65_{0.09}$ | 0,4,6 |
| 1tro | *3* | $2.84_{0.35}$ | $2.07_{0.08}$ | $2.86_{0.23}$ | $2.73_{0.20}$ | $4.56_{0.83}$ | $0.43_{0.03}$ | 0,2,8 |
| 1f4k | *3* | $2.96_{0.11}$ | $2.00_{0.21}$ | $3.05_{0.18}$ | $2.92_{0.18}$ | $3.12_{0.68}$ | $0.40_{0.02}$ | 0,0,10 |
| 1hjc | *2* | $1.83_{0.11}$ | $2.50_{0.22}$ | $2.66_{0.23}$ | $1.94_{0.13}$ | $6.63_{1.52}$ | $0.53_{0.03}$ | 0,5,5 |
| 1vrr | *3* | $3.27_{0.11}$ | $1.89_{0.14}$ | $3.21_{0.10}$ | $2.53_{0.13}$ | $5.51_{1.60}$ | $0.47_{0.02}$ | 0,4,6 |
| 3cro | *3* | $2.34_{0.14}$ | $1.87_{0.12}$ | $2.44_{0.13}$ | $2.10_{0.07}$ | $3.08_{0.84}$ | $0.54_{0.04}$ | 0,0,10 |
| 1r40 | *3* | $2.97_{0.22}$ | $2.46_{0.12}$ | $3.02_{0.18}$ | $3.00_{0.16}$ | $3.68_{1.02}$ | $0.35_{0.05}$ | 0,3,7 |
| 1mnn | *2* | $1.65_{0.03}$ | $1.65_{0.14}$ | $1.96_{0.17}$ | $1.84_{0.11}$ | $4.37_{2.59}$ | $0.52_{0.03}$ | 0,0,10 |
| 1wot | *3* | $3.93_{0.48}$ | $2.49_{0.16}$ | $3.65_{0.60}$ | $3.28_{0.40}$ | $4.57_{2.51}$ | $0.41_{0.05}$ | 0,0,10 |
| 1by4 | *3* | $5.06_{0.59}$ | $1.89_{0.12}$ | $4.62_{0.51}$ | $4.17_{0.21}$ | $5.32_{1.23}$ | $0.37_{0.06}$ | 0,7,3 |
| 1ddn | *5* | $5.39_{1.96}$ | $2.31_{0.17}$ | $5.51_{1.84}$ | $4.99_{1.76}$ | $7.85_{3.68}$ | $0.49_{0.04}$ | 2,5,3 |
| 1k79 | *2* | $1.69_{0.07}$ | $2.63_{0.24}$ | $2.66_{0.37}$ | $2.54_{0.33}$ | $5.68_{2.14}$ | $0.40_{0.09}$ | 0,1,9 |
| 1pt3 | *2* | $2.44_{0.07}$ | $1.54_{0.16}$ | $2.65_{0.19}$ | $2.57_{0.31}$ | $6.50_{1.12}$ | $0.39_{0.07}$ | 0,4,6 |
| 1cma | *3* | $3.27_{0.14}$ | $1.39_{0.08}$ | $3.15_{0.14}$ | $3.13_{0.09}$ | $4.11_{0.52}$ | $0.54_{0.04}$ | 0,1,9 |
| **"intermediate"** | | | | | | | | |
| 1fok | *2* | $1.82_{0.05}$ | $2.42_{0.12}$ | $2.28_{0.07}$ | $2.72_{0.23}$ | $9.28_{1.34}$ | $0.16_{0.04}$ | 0,10,0 |
| 1jj4 | *2* | $4.28_{0.20}$ | $2.31_{0.11}$ | $4.22_{0.24}$ | $3.76_{0.29}$ | $5.58_{1.04}$ | $0.34_{0.03}$ | 0,7,3 |
| 1h9t | *3* | $2.18_{0.13}$ | $2.63_{0.24}$ | $2.55_{0.22}$ | $2.33_{0.34}$ | $6.89_{2.99}$ | $0.46_{0.06}$ | 0,4,6 |
| 2irf | *2* | $4.58_{0.23}$ | $2.20_{0.21}$ | $4.47_{0.34}$ | $4.18_{0.37}$ | $6.60_{1.10}$ | $0.32_{0.03}$ | 0,10,0 |
| 1z63 | *2* | $2.85_{0.04}$ | $3.46_{0.07}$ | $3.33_{0.07}$ | $3.03_{0.09}$ | $7.91_{0.56}$ | $0.54_{0.05}$ | 0,10,0 |
| 7mht | *2* | $3.78_{0.10}$ | $2.91_{0.12}$ | $3.81_{0.11}$ | $6.70_{0.28}$ | $6.25_{1.58}$ | $0.20_{0.04}$ | 0,10,0 |
| 2fl3 | *2* | $4.21_{0.06}$ | $2.29_{0.12}$ | $4.07_{0.11}$ | $5.39_{0.15}$ | $7.09_{1.53}$ | $0.27_{0.05}$ | 1,9,0 |
| 1rva | *2* | $3.67_{0.28}$ | $3.60_{0.25}$ | $3.72_{0.25}$ | $4.09_{0.27}$ | $5.65_{0.86}$ | $0.28_{0.02}$ | 0,9,1 |
| 1eyu | *2* | $3.46_{0.36}$ | $2.73_{0.25}$ | $3.41_{0.37}$ | $3.52_{0.39}$ | $3.52_{0.72}$ | $0.38_{0.05}$ | 0,1,9 |
| 1b3t | *2* | $3.40_{0.07}$ | $2.59_{0.14}$ | $3.36_{0.06}$ | $3.88_{0.07}$ | $5.51_{0.46}$ | $0.31_{0.03}$ | 0,10,0 |
| 1diz | *2* | $1.17_{0.04}$ | $5.04_{0.18}$ | $2.93_{0.17}$ | $2.82_{0.21}$ | $12.02_{3.43}$ | $0.33_{0.04}$ | 0,10,0 |
| 1bdt | *3* | $4.96_{0.24}$ | $1.98_{0.09}$ | $4.46_{0.24}$ | $4.28_{0.24}$ | $6.74_{0.90}$ | $0.28_{0.02}$ | 0,10,0 |
| 1jto | *2* | $6.50_{0.31}$ | $4.37_{0.20}$ | $6.37_{0.29}$ | $5.17_{0.23}$ | $7.77_{1.15}$ | $0.25_{0.02}$ | 0,10,0 |

**Table 3.** Continued

| | | RMSD (Å) | | | | | fnat[f] | CAPRI[g] |
|---|---|---|---|---|---|---|---|---|
| **complex** | | **protein[a]** | **DNA[b]** | **Total[c]** | **Interface[d]** | **ligand[e]** | | **-,*,**** |
| 2fio | *2* | $2.48_{0.18}$ | $4.62_{0.23}$ | $4.55_{0.25}$ | $3.48_{0.19}$ | $4.32_{0.99}$ | $0.26_{0.05}$ | 0,10,0 |
| 1kc6 | *3* | $5.04_{0.47}$ | $3.90_{0.19}$ | $5.06_{0.43}$ | $3.74_{0.29}$ | $13.34_{2.85}$ | $0.30_{0.04}$ | 1,9,0 |
| 1zs4 | *2* | $5.86_{0.15}$ | $3.27_{0.25}$ | $5.93_{0.26}$ | $4.87_{0.22}$ | $10.17_{2.78}$ | $0.23_{0.04}$ | 9,1,0 |
| 1emh | *2* | $2.02_{0.04}$ | $4.10_{0.09}$ | $2.71_{0.13}$ | $2.95_{0.15}$ | $7.81_{2.53}$ | $0.20_{0.03}$ | 0,10,0 |
| **"difficult"** | | | | | | | | |
| 3bam | *3* | $5.77_{0.10}$ | $2.46_{0.12}$ | $5.55_{0.10}$ | $6.94_{0.18}$ | $7.17_{2.65}$ | $0.21_{0.02}$ | 2,8,0 |
| 1a74 | *2* | $1.90_{0.10}$ | $4.43_{0.24}$ | $3.10_{0.13}$ | $3.80_{0.17}$ | $5.44_{1.10}$ | $0.24_{0.02}$ | 0,10,0 |
| 4ktq | *2* | $2.50_{0.10}$ | $2.69_{0.17}$ | $3.21_{0.13}$ | $3.81_{0.18}$ | $16.4_{2.48}$ | $0.22_{0.06}$ | 0,10,0 |
| 1g9z | *2* | $4.56_{0.07}$ | $2.16_{0.15}$ | $4.08_{0.07}$ | $2.76_{0.07}$ | $4.22_{0.18}$ | $0.32_{0.03}$ | 0,2,8 |
| 1zme | *2* | $6.96_{1.09}$ | $3.05_{0.14}$ | $8.22_{2.13}$ | $5.02_{1.02}$ | $8.73_{3.33}$ | $0.11_{0.08}$ | 4,6,0 |
| 1qne | *2* | $1.79_{0.07}$ | $4.17_{0.21}$ | $2.90_{0.10}$ | $3.18_{0.12}$ | $3.29_{0.93}$ | $0.19_{0.02}$ | 0,10,0 |
| 1qrv | *3* | $6.28_{0.93}$ | $4.89_{0.18}$ | $6.36_{0.73}$ | $4.82_{0.22}$ | $4.89_{0.41}$ | $0.28_{0.04}$ | 0,9,1 |
| 1vas | *2* | $1.78_{0.09}$ | $5.09_{0.31}$ | $3.85_{0.30}$ | $3.75_{0.41}$ | $5.75_{1.79}$ | $0.15_{0.06}$ | 2,8,0 |
| 1azp | *2* | $3.37_{0.09}$ | $2.19_{0.20}$ | $3.43_{0.14}$ | $2.53_{0.14}$ | $4.93_{0.49}$ | $0.27_{0.04}$ | 0,8,2 |
| 1z9c | *3* | $7.60_{0.26}$ | $3.38_{0.12}$ | $7.03_{0.62}$ | $7.25_{0.41}$ | $10.37_{3.12}$ | $0.17_{0.05}$ | 3,7,0 |
| 103t | *2* | $3.03_{0.07}$ | $5.06_{0.23}$ | $4.96_{0.39}$ | $4.95_{0.23}$ | $12.67_{2.77}$ | $0.15_{0.02}$ | 2,8,0 |
| 2oaa | *2* | $8.16_{0.16}$ | $3.46_{0.26}$ | $7.86_{0.20}$ | $7.67_{0.27}$ | $16.40_{2.58}$ | $0.23_{0.06}$ | 10,0,0 |
| 1dfm | *3* | $7.64_{0.45}$ | $2.62_{0.23}$ | $7.89_{0.35}$ | $8.15_{0.18}$ | $13.02_{1.24}$ | $0.07_{0.02}$ | 10,0,0 |

Average RMSD values (Å) were calculated after superimposition of all heavy atoms belonging to the protein (a) the DNA (b), the full complex (c) and the interface (d). The ligand RMSD (e) was calculated by superimposition on all phosphate atoms of the DNA and subsequently on all Cα atoms of the proteins. (f) Fraction of native contact and (g) the number of no-, one- and two-star CAPRI solutions. Standard deviations are in subscript. The number of individual components docked is shown in italic after the PDB id.

## Summary

One of the central questions in system biology nowadays is; how do biomolecules interact to perform their function? Detailed knowledge of the structure of biomolecular complexes at atomic resolution provides valuable insights into their function and impacts on many different fields such as medical sciences, molecular biology and pharmacology. Nuclear Magnetic Resonance (NMR) and X-ray crystallography are experimental techniques that are vital in answering this question. Although these methods will remain indispensible to the structural biologist, they have their limitations and as a consequence, there is a large group of complexes that is extremely difficult to solve. Among these are transient, short-lived complexes, membrane associated complexes and protein nucleic-acid complexes, all of which are biologically very interesting.

In the last decade, the toolbox of the structural biologist has been extended with a set of powerful computational techniques, notably computational docking. Docking is the art of modelling a complex in its "bound" state from its "unbound", free components, using a variety of computational algorithms. As such, docking provides an appealing alternative for those cases where experimental techniques encounter difficulties in solving the 3D structure of a complex. Both protein-protein and protein-ligand docking are nowadays commonly used in academia and industry. While docking is flourishing in these fields, less progress has been made in the development of successful protein-DNA docking methods.

DNA-interacting proteins, however, fulfil an important role in the biomolecular interaction networks of the living cell. If changes in the environment of a cell require adaption of its protein content, it is eventually up to the DNA-interacting proteins to induce changes in the level of transcription in the cell. Furthermore, sophisticated protein machineries guard the DNA against hazardous influences that may damage it. A disruption of this delicate balance of regulation can lead to severe consequences for the cell ranging from the inability to perform certain functions to cancer or cell death. An in-depth understanding of the mechanism underlying these regulation processes is only possible by studying them at atomic detail. Although the number of solved protein-DNA complexes in the RCSB protein database (http://www.rcsb.org) is steadily growing, the number of DNA-binding proteins and DNA sequence motifs to which proteins can bind is much larger. As a consequence it becomes infeasible to solve the structures of all complexes using experimental techniques. This is where computational docking can be of assistance. The work described in this thesis focuses on the development of an efficient protein-DNA docking method.

In **chapter 2** the protein-DNA docking field is introduced by an indebt review of its history, the associated docking challenges and the way various methods try to deal with them. The chapter portraits Protein-DNA docking as a young but growing discipline driven by an increased understanding of the vital role these complexes fulfil in the living cell. The field has developed itself along the same lines as protein-protein docking, having to deal with many of the same challenges. The omnipresence of conformational changes and the identification of the DNA interaction interfaces are, however, challenges unique to this discipline. The chapter concludes with the theoretical foundation of a unique two-stage protein-DNA docking method implemented in HADDOCK that aims to deal with these challenges.

**Chapter 3** provides the proof of principle of the two-stage docking method formulated in **chapter 2**. The method was tested using

the three unbound monomeric structures of otherwise dimeric transcription factors and their respective operator half-sides (bacteriophage 434 Cro, phage λ Arc and *Escherichia Coli Lac*). The results illustrate the ability of HADDOCK to drive the docking based on biochemical and biophysical data and successfully reconstruct the correct interface. Haddock was able to induce the specific DNA conformational changes that lead to the final conformation of the DNA in the complex when starting from a canonical B-DNA structural model. However, the full transition from the DNA unbound to bound conformation could not be made without a gross DNA helical deformation. Nevertheless, an analysis of the induced DNA conformational changes showed that the primary bend and twist motions that are at the origin of these changes were modelled. This information was sufficient to introduce DNA bending and twisting in a modelling step using the geometrical relationship between these motions and the base pair step parameters Roll and Twist.
The generated ensemble of custom DNA structural models more closely represented the bound conformation of the DNA in the complex and did not suffer from helical deformations. This ensemble was used as input for a second, "refinement", docking round. The near native docking solutions obtained in this two-stage protocol reproduce many of the molecular contacts and specific DNA conformational changes observed in the experimental structures.

However, the biological relevant, dimeric conformation of the used test cases are considerably more challenging than their monomeric counterparts with respect to their size, interface complexity and DNA conformational changes. A successful application of the two-stage docking method to these systems requires an improved DNA modelling approach and a larger number of test cases to evaluate the methods performance. These are described in **chapters 4** and **5**.

**Chapter 4** describes an improved method for the analysis and modelling of DNA bending called 3D-DART (3DNA driven DNA Analysis and Rebuilding Tool). The method describes DNA bending as an accumulation of bend vectors between successive base pairs relative to a common origin in global Euclidean space. These bend vectors are subsequently transformed into Roll, Tilt and Twist values in the local base pair reference frame. The latter values can be used to model DNA structures with custom bending. The algorithms functionality is made available through an intuitive web interface accessible at http://haddock.chem.uu.nl/dna.

**Chapter 5** describes a dedicated protein-DNA docking benchmark, carefully constructed to provide a representative selection of protein-DNA complexes. The protein and DNA structures of all 47 complexes in the benchmark are present in both their bound and unbound conformation. The benchmark contains a variety of challenging systems in terms of the size of their interaction interface, the number of individual components in the complex and the conformational changes initiated upon complex formation. The variety of test cases in this non-redundant benchmark makes it a useful tool for validation, development and comparison of protein-DNA docking methods.

In **chapter 6** the protein-DNA docking method described in **chapter 3** is challenged using the protein-DNA benchmark (**chapter 5**). The DNA analysis and rebuilding tool described in **chapter 4** is used in the DNA modelling stage of the docking method. The results illustrate the robustness of the method to accurately model the variety of DNA conformational changes present in the benchmark and to assemble the protein-DNA

interfaces using the data-driven approach implemented in HADDOCK. For many of the test cases, the two-stage docking approach was the only way to overcome the limitation of dealing with large conformational changes within a single docking run. Despite the excellent results, there are still a number of test cases that pose a considerable challenge to the method.

Finally, in **chapter 7**, the implications of this work for the protein-DNA docking field in general are discussed. It becomes clear, that only through a community effort, the protein-DNA docking field can mature. With the methods and tools described in this thesis I hope to have set the stage for further method development, stimulate others to take the challenges and leverage protein-DNA docking to the next level of accuracy.

## Samenvatting

Hoe gaan biomoleculen een interactie met elkaar aan om zo hun functie te kunnen vervullen? Dat is één van de centrale vragen in de huidige systeembiologie. De studie van deze biomoleculaire complexen op atomair niveau levert waardevolle en gedetailleerde kennis op over hun functie, die van vitaal belang is voor onder andere de medische wetenschap, moleculaire biologie en farmacie.
Kernspinresonantie, ook wel NMR (Nuclear Magnetic Resonance) genoemd, tezamen met röntgenkristallografie zijn experimentele technieken, die cruciaal zijn bij studie van deze structuren op atomair niveau en het beantwoorden van die centrale vraag. Alhoewel deze technieken altijd belangrijk zullen blijven voor de structuurbioloog hebben zij hun limitaties. Als gevolg hiervan is er een grote groep biomoleculaire complexen die extreem moeilijk te bestuderen is. Hieronder vallen bijvoorbeeld complexen die een zwakke en vaak kortstondige interactie met elkaar hebben, complexen geassocieerd aan biomembranen en eiwit-nucleïnezuur complexen. Al deze complexen zijn echter biologisch erg interessant.
In de laatste tien jaar is de gereedschapskist van de structuurbioloog echter uitgebreid met een collectie krachtige computertechnieken, in het bijzonder docking. Docking is de kunst van het modelleren van een biomoleculair complex in zijn gebonden staat met de structuren van de individuele biomoleculen in hun vrije, niet gebonden staat met behulp van verschillende algoritmen. Deze techniek biedt een aantrekkelijk alternatief bij de studie van biomoleculaire complexen daar waar de traditionele experimentele technieken tekort schieten. Zowel de docking van eiwit met eiwit als eiwit met kleine liganden wordt tegenwoordig veelvuldig gebruikt in de academie en industrie. Alhoewel docking in deze gebieden snel volwassen wordt, lopen de ontwikkelingen in het veld van eiwit-DNA docking achter.
Eiwitten die een interactie aangaan met DNA vervullen echter een belangrijke rol in de biomoleculaire interactie netwerken in de levende cel. Wanneer veranderingen in de leefomgeving van de cel een aanpassing van zijn eiwitsamenstelling verlangt zijn het uiteindelijk de DNA bindende eiwitten die de benodigde wijziging in transcriptie initiëren. Daarnaast is er een breed scala aan eiwitten die samen het DNA beschermen tegen beschadigingen door externe factoren. Wanneer dit delicate proces van regulatie wordt verstoord kan dit nadelige effecten voor de cel tot gevolg hebben, zoals het niet kunnen uitvoeren van bepaalde functies, kanker of uiteindelijk de dood van de cel. Ook het verkrijgen van een gedetailleerd inzicht in dit proces van regulatie kan alleen door de studie van de eiwit-DNA complexen op atomair niveau. Alhoewel het aantal eiwit-DNA complexen in de online RCSB eiwit database (http://www.rcsb.org) gestaag toeneemt, is het totaal aantal DNA bindende eiwitten en eiwit interactie motieven op het DNA vele malen groter. Gelet op het grote aantal mogelijke eiwit-DNA complexen en de tekortkomingen van de experimentele technieken, is het extreem moeilijk al deze complexen middels experimentele technieken op te lossen. Hier kan docking middels een computer een belangrijke rol vervullen.
Het onderzoek beschreven in dit proefschrift richt zich op de ontwikkeling van een efficiënte eiwit-DNA docking methode.

In **Hoofdstuk 2** wordt het eiwit-DNA docking veld geïntroduceerd door middel van een gedetailleerd overzicht van de historie, de uitdagingen waar het veld mee te maken heeft en de wijze waarop de verschillende methoden hier een oplossing op trachten te vinden. Het wordt duidelijk dat het eiwit-DNA dockingveld een jonge maar groeiende discipline is, gedreven door de vitale rol die

deze complexen vervullen in de levende cel. Het veld ontwikkelt zich op dezelfde wijze als in eiwit-eiwit docking en deelt hiermee veel van dezelfde uitdagingen. Toch zijn factoren zoals de alom aanwezige conformationele veranderingen en de identificatie van de DNA interactie interfaces unieke uitdagingen voor dit veld. Het hoofdstuk wordt afgesloten met een beschrijving van de HADDOCK docking methoden en de formulering van het theoretisch fundament van een uniek twee stadia eiwit-DNA docking protocol geïmplementeerd in deze docking methode.

In **hoofdstuk 3** wordt de bovenstaande methode gevalideerd met behulp van drie test complexen bestaande uit monomere transcriptie factoren in hun niet gebonden staat en hun DNA operator (bacteriofaag 434 Cro, phage λ Arc and *Escherichia Coli Lac*). De resultaten laten duidelijk zien dat HADDOCK in staat is de correcte eiwit-DNA interface the modelleren op basis van biochemische en biofysische data. Wanneer wordt gestart met een ideale B-DNA structuur, is de introductie van expliciete flexibiliteit voldoende om de aanzet te geven tot de DNA conformationele veranderingen, die resulteren in de uiteindelijke conformatie van het DNA in het complex. Deze flexibiliteit is echter onvoldoende om de gehele overgang van het DNA van een vrije naar een gebonden conformatie te realiseren zonder ernstige vervorming van de dubbele helix. Een analyse van de geïntroduceerde vormveranderingen in het DNA wezen uit dat de fundamentele vrijheidsgraden, uitgedrukt in draai- en buigbewegingen, correct waren gemoduleerd. Deze worden door het DNA gebruikt om zijn conformatie aan die van het eiwit aan te passen. De geometrische relatie tussen deze vrijheidsgraden en de base paar stap parameters Roll en Twist zijn voldoende om op basis van de docking resultaten nieuwe DNA modellen te genereren. Deze verzameling van modellen komt beter overeen met de conformatie van het DNA in het referentie complex en is vrij van deformaties. De modellen dienen als startmodellen voor een tweede docking ronde waarin de conformatie wordt verfijnd en de uiteindelijke selectie plaatsvindt. De uiteindelijke docking modellen komen sterk overeen met de referentiestructuren. Veel van de inter-moleculaire contacten alsmede de specifieke conformatie van het DNA waren correct gemodelleerd.

Deze initiële resultaten waren veelbelovend maar zijn verkregen met de monomere vorm van drie transcriptiefactoren die in hun biologisch relevante conformatie als dimeer aanwezig zijn. Deze zijn echter vele malen moeilijker te modelleren door hun grootte, de complexiteit van de interface en de vormveranderingen in het DNA. Om ook hier het twee stadia eiwit-DNA docking protocol succesvol te kunnen toepassen dienen er aanpassingen gemaakt te worden en moet het aantal testcomplexen worden uitgebreid. Beide facetten worden beschreven in **hoofdstuk 4** en **5**.

**Hoofdstuk 4** beschrijft 3D-DART (3DNA driven DNA Analysis and Rebuilding Tool), een verbeterde methode voor de analyse en modellering van buigingen in de DNA helix. De methode beschrijft buigen als een accumulatie van buigvectoren tussen opeenvolgende basenparen ten opzichte van een gemeenschappelijke oorsprong in het globale Euclidiaanse referentiekader. Deze buigvectoren worden vervolgens omgezet in Roll, Tilt en Twist basenpaar-stap parameters ten opzichte van het lokale basenpaar referentiekader. Deze parameters kunnen gebruikt worden voor de generatie van DNA modellen en geven als zodanig volledige controle over de wijze waarop het DNA buigt. De functionaliteit van het algoritme is beschikbaar gemaakt via een intuïtieve webserver (http://haddock.chem.uu.nl/dna).

In **hoofdstuk 5** wordt een specifieke eiwit-DNA docking benchmark beschreven, zorgvuldig samengesteld met een representatieve selectie complexen. De eiwit- en DNA structuren van alle 47 complexen zijn aanwezig in zowel hun gebonden als ongebonden conformatie. De benchmark bevat een brede variëteit aan uitdagende complexen gelet op hun grootte, het aantal individuele componenten waaruit de complexen bestaan en de conformationele veranderingen die zij ondergaan wanneer zij het complex vormen. De variëteit in deze non-redundant benchmark maakt het een geschikt gereedschap voor de validatie, ontwikkeling en vergelijking van de verschillende dockingmethoden.

In **hoofdstuk 6** wordt de initiële twee-stadia eiwit-DNA docking methode toegepast op de complexen in de benchmark van **hoofdstuk 5.** De DNA analyse en moduleringsstap in de methode wordt nu uitgevoerd door middel van het algoritme beschreven in **hoofdstuk 4**. De resultaten onderstrepen het robuuste karakter van de methode in het modelleren van de grote variëteit aan DNA conformationele veranderingen en reconstructie van de interfaces middels datagedreven docking. Voor veel complexen was het twee-stadia docking proces de enige manier om succesvol om te gaan met de grote conformationele veranderingen in het DNA welke niet in één HADDOCK run geïntroduceerd konden worden. Ondanks de uitstekende resultaten zijn er nog steeds een aantal testsystemen die een uitdaging vormen.

Tot slot worden in **hoofdstuk 7** de implicaties van het onderzoek beschreven in dit proefschrift voor het eiwit-DNA docking veld in het algemeen besproken. Hierbij is duidelijk dat een verdere ontwikkeling alleen mogelijk is door een gemeenschappelijke aanpak. Ik hoop dan ook dat dit proefschrift hiervoor de basis zal vormen, zodat een samenwerking van de eiwit-DNA docking gemeenschap het veld op een volgend niveau kan brengen.

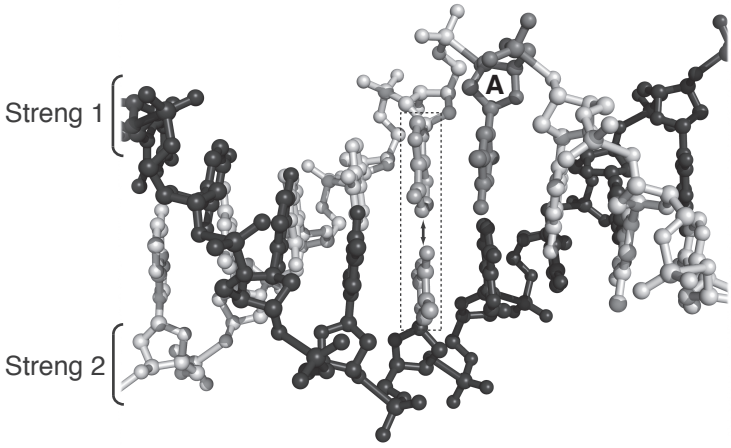
## Samenvatting vereenvoudigd

### DNA, de blauwdruk van het leven

U draagt, net als de mensen om u heen, een "blauwdruk" met u mee waarin beschreven staat hoe u in elkaar steekt, tenminste, fysiek dan. Deze blauwdruk wordt ook wel het DNA genoemd, en is de belangrijkste drager van erfelijke informatie in alle bekende organismen (box 1). Het DNA bevat de informatie die nodig is om de grote verscheidenheid aan eiwitten te bouwen die de cellen in uw lichaam rijk zijn. Elke cel heeft zijn eigen kopie van deze blauwdruk en weet, in principe, hoe het hele lichaam waarvan het deel uitmaakt, is opgebouwd. Dat is veel meer informatie dan de cel in zijn leven nodig heeft. Een cel die deel uitmaakt van de hersenen hoeft zich alleen maar bezig te houden met de rol die het daar vervult en heeft niets aan de informatie die vertelt hoe het een spiercel moet zijn. Het is dus belangrijk dat de cel precies op de juiste tijd en de juiste plek in het lichaam doet wat het moet doen en dat de informatie altijd in de juiste vorm beschikbaar is. Dit is de taak van

**Box 1.**

DNA (Desoxyribonucleïnezuur, figuur 1) is één lang molecuul, wel twee meter in elke cel van het menselijk lichaam, dat bestaat uit twee strengen van aan elkaar gekoppelde nucleotiden. Die twee strengen samen buigen tot een dubbele helix. Elk van de vele nucleotiden in de twee strengen bevat één base waarvan er slechts vier typen zijn: adenine (A), thymine (T), guanine (G) en cytosine (C). De basen van twee tegenover elkaar liggende nucleotiden kunnen een interactie met elkaar aangaan en vormen een basenpaar. De twee nucleotidestrengen zijn met elkaar verbonden via de vele basenparen. Er zijn slechts twee mogelijke basenparen: een adenine paart altijd met een thymine en een guanine altijd met een cytosine. De volgorde van nucleotiden in een streng wordt een sequentie genoemd. Omdat er zeer veel sequenties mogelijk zijn, kan de volgorde van nucleotiden unieke erfelijke informatie verschaffen, de genetische code. Aan de hand van de genetische code kan een specifieke DNA-sequentie vertaald worden in de aminozuursequentie van een eiwit.




**Figuur 1.** Een weergave van de atomaire structuur van een klein stukje DNA van negen nucleotiden lang. Atomen zijn weergegeven als kleine bolletjes, onderling met elkaar verbonden via staafjes. De twee strengen nucleotiden (lichtgrijs en donkergrijs) draaien als een dubbele helix in elkaar. **A**) geeft één nucleotide weer. Het basedeel van twee tegenover elkaar liggende nucleotiden (gestreepte box) hebben een interactie met elkaar die de twee strengen bij elkaar houdt.

eiwitten.

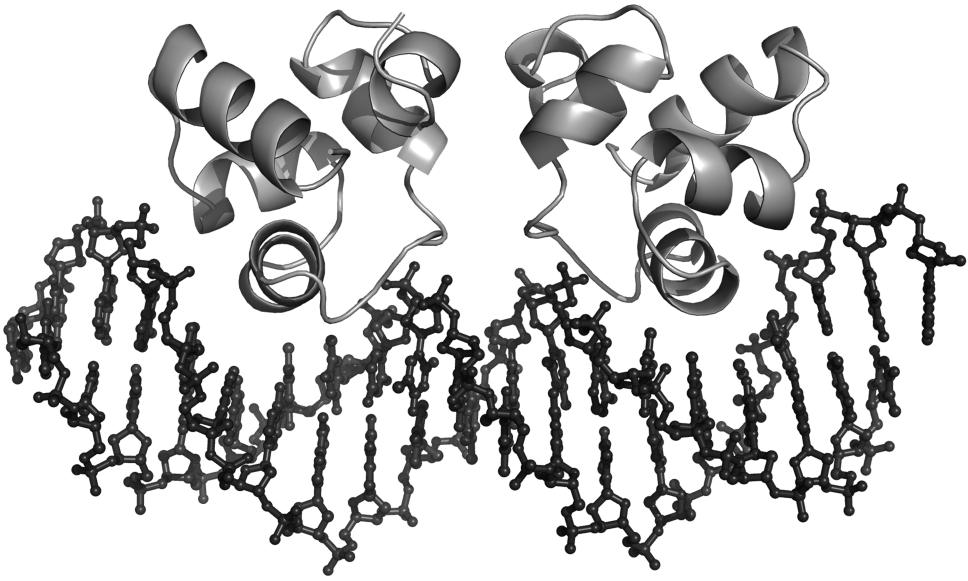
**Eiwitten, de machines van de cel**

Eiwitten zijn de "werkpaarden" van de cel en zorgen ervoor dat deze zijn functie kan vervullen. Er zijn vele duizenden verschillende eiwitten die allemaal een andere taak hebben. Er zijn eiwitten die boodschappen doorgeven, als bouwsteen dienst doen, stoffen transporteren, afbreken of juist opbouwen. Alleen kunnen eiwitten niet veel, maar door aan elkaar te binden vervullen ze hun functie en samen vormen ze een groot netwerk van eiwitten die met elkaar praten. Het DNA speelt ook in deze netwerken een belangrijke rol. Wanneer de cel te weinig of te veel van een bepaald eiwit heeft, dan moet de balans hersteld worden. De cel doet dit door het stukje DNA wat de informatie voor het eiwit bevat, meer beschikbaar te maken of juist te onderdrukken. Dit wordt geregeld door een grote groep eiwitten die affiniteit hebben met het DNA en eraan kunnen binden. Zij kunnen de DNA "code" lezen en zorgen ervoor dat de informatie op het juiste moment beschikbaar komt. Daarnaast zijn zij in staat om het DNA te beschermen tegen kwade invloeden van buitenaf en eventueel te repareren mocht het toch beschadigd raken. Eiwitten en DNA vormen dus een onafscheidelijk paar.

Wanneer deze eiwitten niet goed meer hun functie kunnen vervullen kan dit nare gevolgen hebben zoals ontwikkelingsziekten en kanker. Omdat de interactie tussen eiwit en DNA zo belangrijk is, wil de wetenschap graag weten hoe een eiwit aan het DNA bindt op het meest gedetailleerde niveau, zijn atomaire structuur.

**De details maken het verschil**

Door vele experimenten en observaties neemt de wetenschap aan dat alle stoffen opgebouwd zijn uit atomen, zo ook de eiwitten en het DNA. Eén of meer eiwitten die aan het DNA binden worden samen een



**Figuur 2.** Een weergave van de atomaire structuur van een complex tussen DNA en twee eiwitten. Het DNA (donkergrijs) is weergegeven met atomen als kleine bolletjes onderling met elkaar verbonden via staafjes. De twee eiwitten zijn ook opgebouwd uit atomen maar die zijn hier niet weergegeven. In plaats daarvan is gekozen voor een "cartoon" weergave die goed de ruimtelijke organisatie van het eiwit laat zien.

eiwit-DNA complex genoemd. De tak van de wetenschap die zich bezighoudt met het bestuderen van eiwit en DNA moleculen, op atomair niveau, wordt de structuurbiologie genoemd. Hoe zo'n atomaire structuur er in het echt precies uitziet is moeilijk te zeggen en daarom gebruiken wetenschappers vaak een weergave die een benadering is van de manier waarop atomen in de ruimte zijn gerangschikt. In figuur 1 heeft u dit voor DNA al kunnen zien en figuur 2 laat dit nogmaals voor een eiwit-DNA complex zien. De structuur levert waardevolle informatie op over hoe het eiwit zijn functie vervult of dat niet meer kan als het beschadigd is. De farmaceutische industrie kan deze informatie bijvoorbeeld gebruiken om gericht een nieuw medicijn te ontwikkelen.
Structuurbiologen gebruiken doorgaans twee experimentele technieken om de atomaire structuur van deze complexen "op te lossen"; kernspinresonantie (NMR in het Engels) en röntgenkristallografie. Het oplossen van de structuur met deze technieken is echter niet altijd zo eenvoudig. Veel eiwitten binden slechts kort aan het DNA en het aantal mogelijke eiwit-DNA complexen is zo groot dat het lang zal duren voordat ze allemaal op deze wijze opgelost zijn. Om deze redenen proberen wetenschappers andere manieren te vinden om de complexen alsnog goed te kunnen bestuderen. Eén van deze manieren is het gebruik van computerprogramma's om de losse eiwitten en het DNA in elkaar te puzzelen. Dit puzzelproces wordt docking genoemd. Dat ook dit niet eenvoudig is bewijst wel het beperkte aanbod aan efficiënte eiwit-DNA docking programma's.

**Puzzelen met flexibele puzzelstukjes**

In hoofdstuk 2 van dit proefschrift zet ik alle eiwit-DNA docking programma's, die tot nu toe zijn ontwikkeld, op een rijtje. Ik kijk hierbij naar hun sterke en zwakke punten bij het oplossen van de puzzel en in hoeverre het eindproduct overeenkomt met de karakteristieke eigenschappen van eiwit-DNA complexen. Zo'n complex valt wel te vergelijken met boutjes en moertjes zoals het omslag van dit proefschrift al laat zien. Het DNA lijkt hierbij wat op een bout, het is ook een cilinder en heeft groeven. De (halve) moer is het eiwit. Bout en moer passen precies op elkaar als de groeven in de moer hetzelfde zijn als het draad op de bout. Op eenzelfde wijze passen ook het eiwit en DNA precies in elkaar wanneer zij aan elkaar binden. Met behulp van deze vergelijking valt duidelijk uit te leggen wat de twee grote struikelblokken zijn die het voor de "dockers" moeilijk maakt om de eiwit-DNA puzzels op te lossen:

*Probleem 1*

Het DNA is als het ware een hele lange bout. Bij een echte bout en moer maakt het voor de moer niet zoveel uit waar het zich op de bout bevindt, zolang het maar past. Voor een eiwit op het DNA is dit wel belangrijk. Een eiwit bindt vaak specifiek met een kleine regio (sequentie) op het lange DNA. Voor een dockingprogramma is het vaak niet eenvoudig om de exacte locatie op het DNA te vinden, net als het voor de moer moeilijk is om te weten waar het zich op de bout bevindt. Nu docken wij doorgaans geen eiwit op DNA van miljoenen basenparen lang, maar zelfs voor een kort stukje DNA van enkele tientallen basenparen is dit proces nog steeds moeilijk.

*Probleem 2*

Wanneer de exacte locatie gevonden is moet de moer op de bout gepast worden. Als de vorm van de bout bekend is, is er maar één moer die daar precies op past. Bij eiwit en DNA ligt dat wat anders, zij kunnen van vorm veranderen. Op het omslag valt dit te zien als gebogen bouten en de twee halve moertjes die zich om de bout heen klappen. Wanneer de puzzelstukjes van vorm veranderen en je van tevoren niet precies weet hoe dit gebeurt, wordt het dockingproces ineens vele malen moeilijker.

Tijdens mijn promotieonderzoek heb ik geprobeerd het dockingprogramma HADDOCK zo uit te breiden dat het goed overweg kan met deze twee problemen. HADDOCK is ontwikkeld in ons laboratorium en is in de loop van de jaren uitgegroeid tot een populair dockingprogramma dat door veel wetenschappers wereldwijd wordt gebruikt. HADDOCK heeft twee grote voordelen die het een veelbelovende methode maakt om eiwitten op DNA te docken. Allereerst stelt HADDOCK de wetenschapper in staat om alle informatie die over de interactie tussen eiwit en DNA vergaard is te gebruiken om te docken. Hierdoor kan beter bepaald worden waar op de bout de moer zal binden en hoe hij dat doet. Die specifieke locatie noemen wij de interactie interface. Ten tweede stelt HADDOCK de eiwitten en het DNA in staat om van vorm te veranderen tijdens het docken om zo beter aan elkaar te binden. Dit laatste is heel belangrijk voor deze complexen maar is tegelijkertijd heel lastig om goed te voorspellen. Hoe groter de veranderingen zijn hoe groter de kans is dat we ernaast zitten en de structuur zijn juiste vorm verliest. Vooral het DNA is hier vatbaar voor en kan al snel zijn juiste, dubbele helixstructuur, verliezen. Eigenlijk kan HADDOCK maar in beperkte mate overweg met vormverandering en dat is vaak niet genoeg om het juiste complex te voorspellen, zeker niet voor het DNA. Om dit probleem op te lossen heb ik HADDOCK uitgebreid. Het dockingproces verloopt nu in twee stappen:

*Stap 1*
Tijdens de eerste stap worden de eiwitten en DNA in elkaar gepuzzeld met behulp van experimentele informatie. Ik gebruik hier een recht stukje DNA, omdat ik nog niet weet hoe het van vorm zal veranderen, maar tijdens het dockingproces mogen zowel het eiwit als het DNA wel van vorm veranderen.

*Stap 2*
Na afloop van stap 1 kijk ik goed naar de wijze waarop het DNA van vorm verandert onder invloed van het eiwit. Globaal doet het DNA dat op twee manieren; het buigt net als de bouten op de voorkant van dit proefschrift; en het draait in- en uit elkaar waarbij de groeven in de bout wijder of nauwer worden. Wanneer ik deze buig- en draaibewegingen aantref, en ze zijn consistent, dan kan ik deze informatie gebruiken om nieuwe DNA modellen te maken die al gebogen en gedraaid zijn. Het voordeel hiervan is dat de nieuwe modellen een keurige dubbele helixstructuur hebben, ook al zijn ze al gebogen en gedraaid. Door voorgevormde DNA modellen te gebruiken kan ik meer vormverandering in het DNA introduceren dan waartoe HADDOCK in staat is, zonder dat het DNA zijn dubbele helixstructuur verliest. Op deze manier hoop ik de vorm die het DNA, in het complex heeft, beter te kunnen benaderen. Stap 2 wordt afgerond met een tweede HADDOCK puzzelronde waar ik de nieuwe DNA modellen gebruik om het uiteindelijke complex te vormen.

In **hoofdstuk 3** test ik de twee stappen methode voor het eerst op drie eenvoudige eiwit-DNA complexen. Deze bestaan elk uit één eiwit dat bindt aan een klein stukje DNA. De structuur van deze testcomplexen is al opgelost met röntgenkristallografie. Ik kan mijn dockingresultaten daarom vergelijken met het echte complex om te zien hoe goed de methode presteert.
De resultaten laten zien dat de methode goed in staat is om de juiste interface te reconstrueren en de vormveranderingen in het eiwit en DNA te introduceren die nodig zijn om het uiteindelijke complex te vormen. Met andere woorden: de dockingresultaten zijn bijna niet van het echte complex te onderscheiden.
De vormveranderingen in het DNA voor deze testcomplexen zijn echter klein. Het kleine stukje DNA buigt vaak maar op één

vaste manier, waarbij de groef iets wijder wordt. In veel eiwit-DNA complexen zijn de vormveranderingen echter veel groter en ingewikkelder. Vaak binden er twee of zelfs meer eiwitten naast elkaar op het DNA, zoals te zien is in figuur 2 en de twee moertjes naast elkaar op één bout op de voorkant van dit proefschrift. De vormveranderingen in het DNA worden nu veel groter waarbij het DNA bijvoorbeeld niet meer op een uniforme wijze buigt. Om ook dit soort complexen goed te kunnen docken heb ik de twee stadia methode verder uitgebreid.

In **hoofdstuk 4** beschrijf ik 3D-DART, een softwareprogramma dat ik heb ontwikkeld om DNA modellen te maken. 3D-DART wordt aangeboden via het internet (http://haddock.chem.uu.nl/dna) waar de gebruiker op een eenvoudige wijze modellen kan maken met volledige controle over de manier waarop het DNA buigt. Deze software is een belangrijke aanvulling op stap 2 van het dockingproces.

Het is belangrijk dat de uitgebreide methode ook getest kan worden op een grote verscheidenheid aan verschillende eiwit-DNA complexen. Ik heb hiervoor speciaal een set van 47 verschillende eiwit-DNA complexen samengesteld. In **hoofdstuk 5** beschrijf ik deze complexen, waarvan de structuur al is opgelost. Al deze complexen verschillen van elkaar in het aantal eiwitten dat bindt op het DNA en de mate waarin eiwit en DNA van vorm veranderen. Aan de hand van deze verschillen heb ik de complexen in drie groepen ingedeeld die de dockingmoeilijkheidsgraad aangeeft, van makkelijk tot moeilijk.

Tot slot test ik de methode op de 47 eiwit-DNA testsystemen in **hoofdstuk 6**. Na het docken van al deze complexen bleek dat deze twee stappen HADDOCK methode goed in staat is de vaak grote vormveranderingen te voorspellen. Hiermee is deze methode tevens de eerste die op zo'n grote schaal overweg kan met vormveranderingen in eiwit-DNA complexen.

## Acknowledgments

It is the 23rd of December 2009 in a small but cosy apartment on Ameland, one of the five islands in the north of the Netherlands. The weather is harsh, it is foggy and cold and the wind is howling around the building.
With a fresh and hot cup of coffee I started writing the last, and in some respect, most important words of this thesis, the acknowledgments.
It is time to look back, and I quickly conclude that the last four years have past all too fast. It seems like yesterday that I started my minor research project at the NMR spectroscopy department. Starting a research project in the field of bioinformatics and structural biology was a bit of a gamble for me. I never worked in an Unix computer environment before and programming was alien to me. I also have to admit that my knowledge of biochemistry wasn't all that thorough either. But according to Alexandre, my supervisor, that all did not matter. Ok... I guess.

He was right, I enjoyed this new discipline and the project was fruitful. Still, I was surprised to got offered a position as a PhD student allowing me to continue the project.
Here, I have to make another confession; I never planned to pursue a career in science and continue as a PhD student after my graduation as a molecular biologist. Nevertheless I accepted the offer because I suspected that a PhD research project would be a valuable experience and provide me with the ability to learn new treats that could be valuable in a future career whatever that me be. Now, four years down the road, I must say that those initial expectations were even more valued than I could have imagined. I learned many new treats and was given the opportunity to develop abilities I never knew I was capable of. For that I'm very grateful. Of course, this could never happen if it wasn't for the stimulating environment that the NMR department is, an environment created by both it's scientific and non-scientific staff and guests. To all those colleagues, past and present, I owe big thanks.

First of all I want to thank Alexandre, Rolf and Rob for the offer to continue as a PhD student. A special thanks goes to Alexandre, your continuous enthusiasm and involvement with your students makes a world of difference, I'm grateful that you have been my supervisor.
A thanks goes to Aalt-Jan van Dijk, my day-to-day supervisor during my minor research project and PhD colleague in the years after. I hope we keep in touch.
A special thanks to the staff of the NMR department, past and present; Johan, Hans, Gert, Barbara, Rainer, Albert, Michiel, Rob and Mark. Thanks for all the support both in a professional setting and a social perspective. Without your ongoing efforts, all the facilities we as students take for granted cannot exist. Gert, sorry I did not spent more time in the wet-lab!

Colleagues have come and gone in the last years. Thanks to those colleagues from the first hour; Devashish, Julija, Anding, Jenny, Suat, Hugo, Ludovic, Karine, Kostas, Aurelien, Jeff, Tammo, Gloria, Eiso, Monique, Manuel, Nathalie, Klaartje, Eugene, Roberto and Aart. Many of you have now spread to many exotic locations around the world. Thanks, all the best and I hope to see you again somewhere in the future.
The department has seen several foreign guests over the years, a special thanks to; Victor Hsu for your help with my first article and 24 presence in the lab. Luca, NMR is tough and you know all about that. Nevertheless, with your positive spirit there is always a bright light at the end of the tunnel. Good luck with finishing your manuscript and I hope to see you again in the near future.

The bioinformatics group has grown in the last years and that is in part due to the enthusiasm of a number of colleagues; Sjoerd, office mate and coding guru thanks for all the stimulating conversations we had, your social involvement in the lab and your continuous professional support to all in need. Your bright mind should stay in academia and I'm sure that the world will learn more from you, all the best. Josine, keep an eye on him.
Mickaël, (french) patriot, I'm missing your lively presence already. Wherever your career will take you, be sure that we will visit you (especially if you are going to Canada or Australia). All the best and know that it is never too late to switch to the big Apple!
If it wasn't Mickaël keeping up the spirit than for sure Adrien made sure that a lively atmosphere was guaranteed. Keep practicing your Dutch. MD will never be the same thanks to Tsjerk. Good luck with your future career and maybe we will work together once more in the not so distant future
Thanks to the bunch of bioinformaticians that reinforced the group in the last year; Ezgi, Panagiotis good luck in the years to come. Gijs, how is that bed coming along?

I would like to thank the CcpN-NMR group for their help in coding the CCPN HADDOCK interface and the good times in many of the bars in Cambridge. Wim, Tim, Rasmus, Wayne, Chris thanks a lot.
I would like to thank a number of people that provided professional support at several stages of the research project. Thanks to Marc Parisien from the University of Montreal, Canada for his work on the scripts for the calculation of DNA base-pair and base-pair step deformation energies. The group of Janusz M. Bujnicki (International Institute of Molecular and Cell Biology, Warsaw, Poland) for providing helpful feedback while testing the 3D-DART web server.
The research described in this thesis has been made possible with the financial support provided by the European Community (FP6 STREP project "ExtendNMR", contract no. LSHG-CT-2005-018988, FP6 I3 project "EU-NMR", contract no. RII3-026145 and FP7 I3 project "eNMR", contract no. 213010-e-NMR), the Netherlands Organisation or Scientific research (NWO/CW grant B81-752 and NWO-TOP grant 700.52.303) and from a VICI grant from the Netherlands Organization for Scientific Research (NWO) to A.M.J.J. Bonvin (grant no. 700.96.442).

Last but not least I'm thankful to Cécile, my soulmate. Your language writing skills made sure that my, sometimes poor, language constructions did not reach the public. The great time we spent together on a daily basis made sure that I was not fully consumed by my work. Thanks for being there.

## Publications


- **M. van Dijk**, A.D.J. van Dijk, V. Hsu, R. Boelens and A.M.J.J. Bonvin (2006), "Information-driven Protein-DNA Docking using HADDOCK: it is a matter of flexibility." *Nucl. Acids Res.*, 34: 3317-3325
- S.J. de Vries, A.D.J. van Dijk, M. Krzeminski, **M. van Dijk**, A. Thureau, V. Hsu, T. Wassenaar and A.M.J.J. Bonvin (2007), "HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets." *Proteins*, 69 (4): 726-733
- S.J. de Vries, **M. van Dijk** and A.M.J.J. Bonvin (2008), "The Prediction of Macromolecular Complexes by Docking" in "Prediction of Protein Structures, Functions and Interactions", J.M. Bujnicki (editor), John Wiley and Sons, Ltd, Chichester. United Kingdom.
- **M. van Dijk** and A.M.J.J. Bonvin (2008), "A protein-DNA benchmark", *Nucl. Acids Res.*, 36 (14, online publication): 1-5
- **M. van Dijk** and A.M.J.J. Bonvin (2009), "3D-DART: a DNA structure modelling sever", *Nucl. Acids Res.*, 1 (37, web server issue): 235-239
- **M. van Dijk** and A.M.J.J. Bonvin (2009), "Protein-DNA docking: The Tricks of an Emerging Trade", *Nucl. Acids Res* (submitted for publication).
- S.J. de Vries, **M. van Dijk** and A.M.J.J. Bonvin (2009), "The HADDOCK web server for data-driven biomolecular docking", *Nature Protocols* (submitted for publication).
- D. Das, G. Folkers, **M. van Dijk**, N. Jaspers, J. Hoeijmakers, R. Kaptein and R. Boelens (2009), "The structure of the XPF -ssDNA complex underscores the distinct roles of the XPF and ERCC1 Helix-hairpin-Helix domains in ss/ds DNA recognition", *Nucl. Acids Res.* (submitted for publication).
- **M. van Dijk** and A.M.J.J. Bonvin (2009), "Pushing the limits of what is achievable in protein-DNA docking. Benchmarking the performance of HADDOCK", *Nucl. Acids Res.* (submitted for publication)

## Curriculum Vitae

Marc van Dijk is op 5 december 1977 geboren in Gouda als zoon van Ingrit en Stan van Dijk. In 1995 behaalde hij zijn MAVO diploma aan het Maarten Luther MAVO te Gouda. In datzelfde jaar begon hij aan de MBO opleiding tot laborant in de technische microbiologie aan het Reynevelt College te Delft. Het laatse jaar van deze opleiding volbracht hij op de vakgroep van prof. dr. M. Jetten van het Kluyver Laboratorium aan de Technische Universiteit te Delft in het kader van zijn hoofdvakstage. Het MBO diploma werd behaald in 1999.
Van 1999 to 2000 doorliep hij met goed gevolg het HLO propedeusetraject van de Hogeschool Rotterdam om vervolgens in 2000 te starten met de studie Biologie aan de Universiteit Utrecht. De laatste twee jaar van de studie volgde hij de Masterstudie "Biomolecular Sciences". Zijn hoofdvakstage tijdens deze master vervulde hij bij de microbiologievakgroep onder leiding van prof. dr. H.A.B Wösten en zijn bijvakstage aan de NMR spectroscopie vakgroep onder leiding van dr. Alexandre .M.J.J. Bonvin. Het doctoraal examen Biologie werd op april 2005 behaald.
Van oktober 2005 tot oktober 2009 was hij werkzaam als junior onderzoeker bij de NMR spectroscopie vakgroep van de Universiteit Utrecht onder begeleiding van Prof. dr. Alexandre M.J.J. Bonvin. Dit onderzoek leidde tot het proefschrift wat u nu voor zich heeft.

# Full colour figures

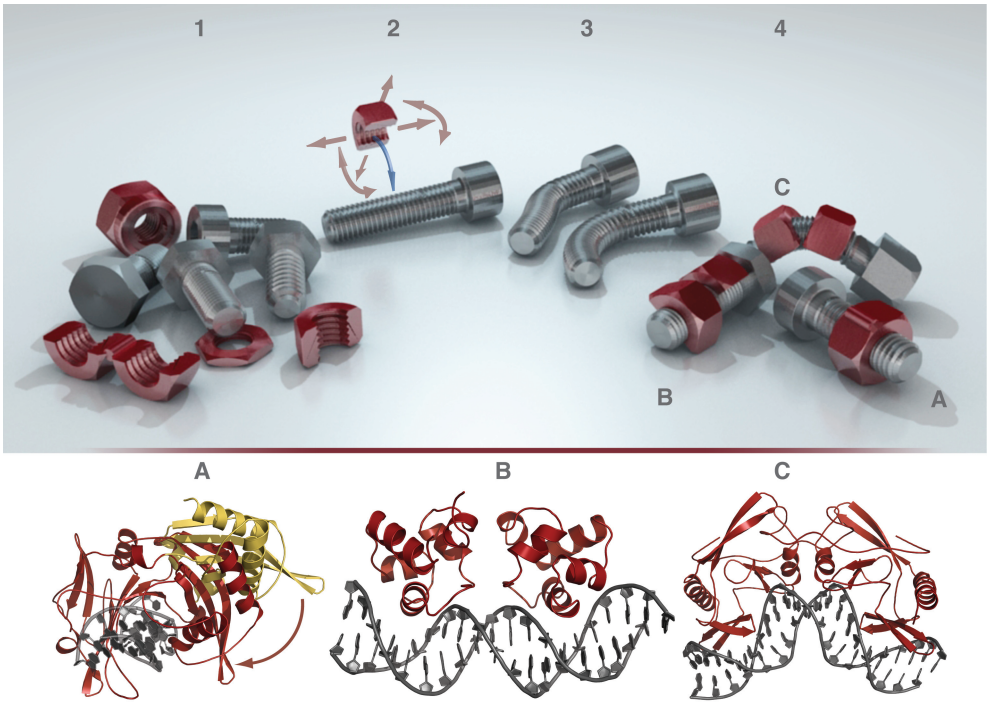



**Figure 2.1.** Schematic representation of the main challenges in protein-DNA docking; **1**) Selecting starting structures for the protein (red bolts) and the DNA (grey screws), which can exist in multiple conformations. **2**) Searching conformational space aimed at assembling the interaction interfaces by rigid body translations and rotations (red arrows). **3**) Dealing with conformational change during the conformational search (DNA, deformed screws). **4**) Ranking the solutions and selecting the "correct" solutions among the many different poses. Examples of complexes: **A**) restriction endonuclease MVAI (2oaa) that undergoes a hinge motion upon DNA interaction from the open state (yellow) to the closed state (red); **B**) Dimeric bacteriophage 434 Cro transcription factor (3cro); and **C**) I-PpoI homing endonuclease (1a74), heavily kinks DNA upon complex formation. The schematic figure in the top panel was generated using Blender (www.blender.org) and the figures in the bottom panel were generated using Pymol (DeLano Scientific, www.pymol.org).



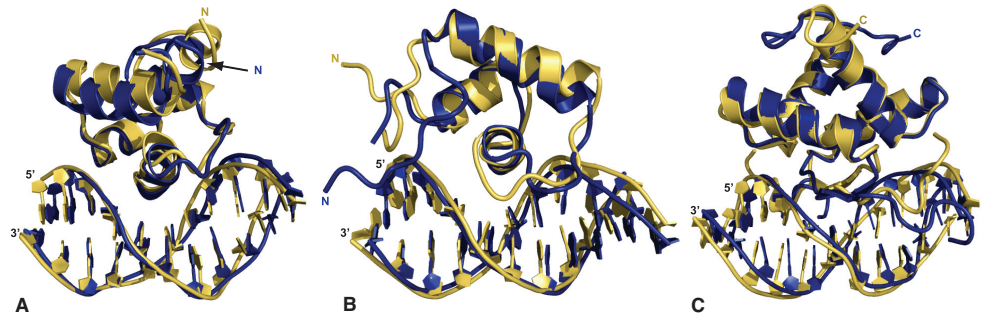



**Figure 3.3.** Best solutions of the unbound flexible docking using a library of pre-bent and twisted DNA structures (blue) superimposed on the reference structure (yellow): Cro-O1R (**A**), Lac-O1 (**B**), Arc-operator (**C**). The structures were superimposed on all heavy atoms of the interface residues (Interface RMSD values: Cro, 1.62 Å; Lac, 2.02 Å; Arc, 1.90 Å). The figures were generated using Pymol (DeLano Scientific LLC, www.pymol.org).

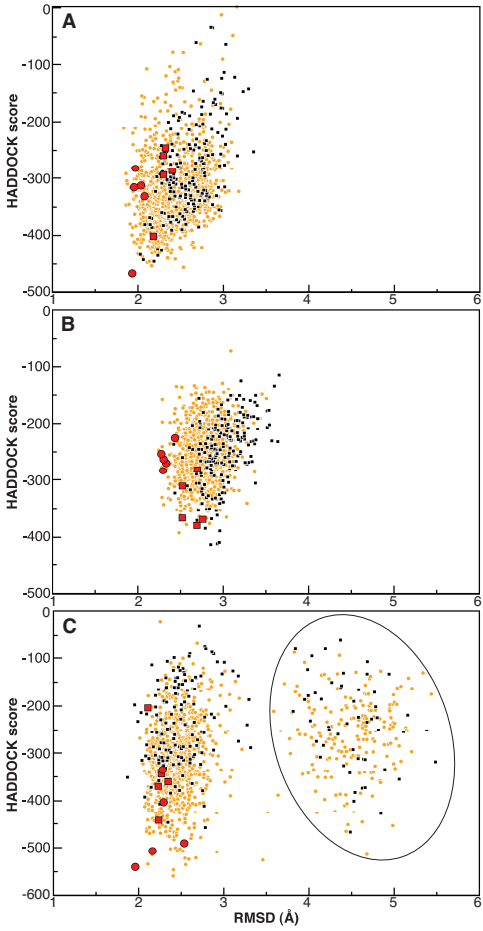



**Figure 3.1.** HADDOCK score versus RMSD from the target (all heavy atoms of the complex) for the Cro (**A**), Lac (**B**) and Arc (**C**) repressors in complex with their operator. Solutions of the unbound flexible docking with canonical B-DNA are shown as small black squares with the five top ranking solutions identified by red squares. Solutions from the docking using a library of pre-bent and twisted DNA structures are shown as small orange circles with the top five ranking solutions identified by red circles. False positives for Arc are shown within a solid ellipse: These correspond to solutions in which the repressor is shifted one or two base pairs along the DNA.

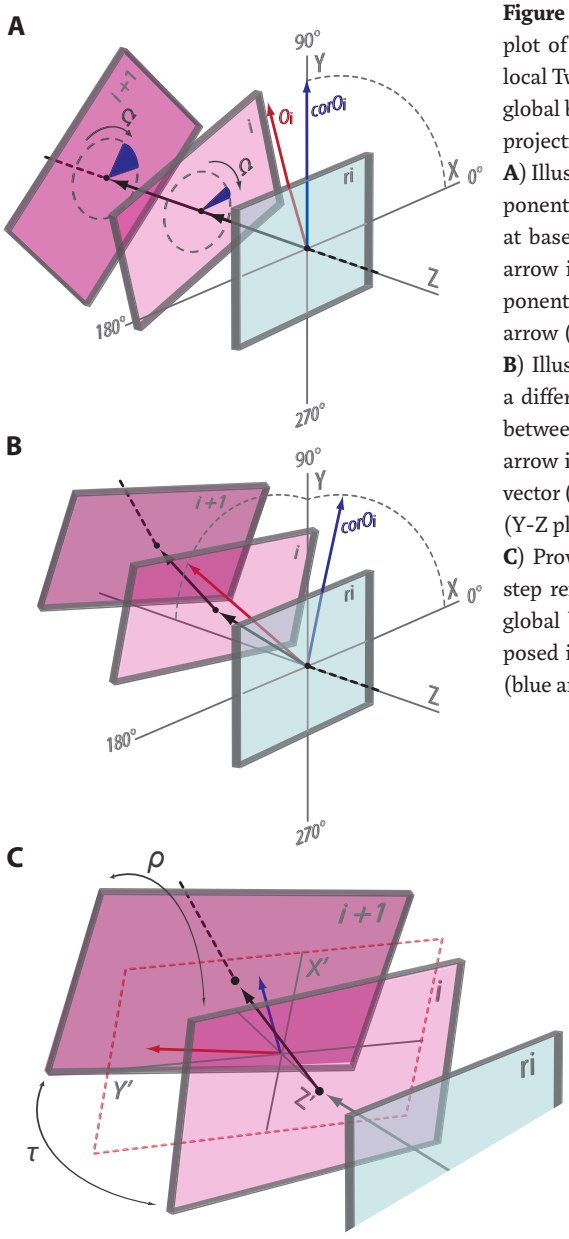



**Figure 4.1.** One block per base-pair Calladin-Drew plot of DNA illustrating the relation between the local Twist (Ω), Roll (ρ) and Tilt (τ) values and the global bend angle for a given base-pair step. Vector projections are normalized for illustrative purposes. **A**) Illustrates the correction of the orientation component of the bend vector for the local Twist value at base pair i and i+1 (blue circle parts). The red arrow indicates the value of the orientation component (Oi) before Twist correction and the blue arrow (corOi aligned with Y-axis) after correction. **B**) Illustrates a bend in the structure as a result of a different bend angle vector (thick black arrows) between every successive base-pair step. The blue arrow illustrates the orientation component of the vector (Y-X plane) and the red arrow the magnitude (Y-Z plane).

**C**) Provides a detailed view or the local base-pair step reference frame between base i and i+1. The global bend vector (thick black arrow) is decomposed into a Tilt (red arrow, Y'-Z' plane) and Roll (blue arrow, X'-Z' plane) contribution.

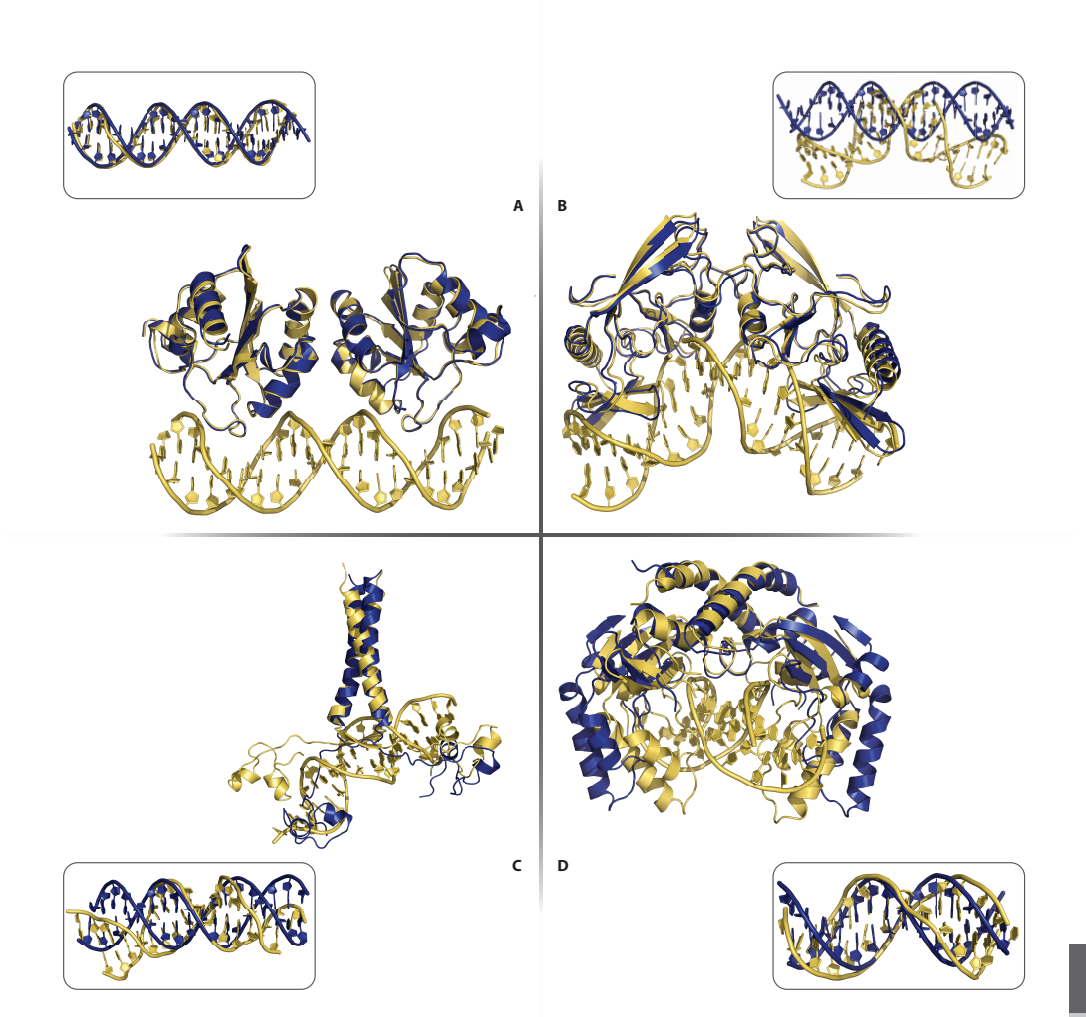
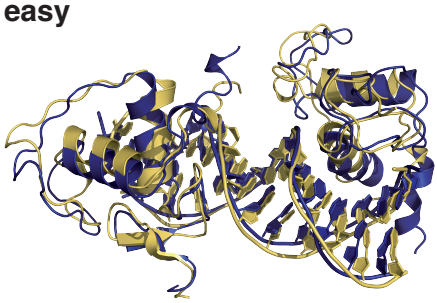


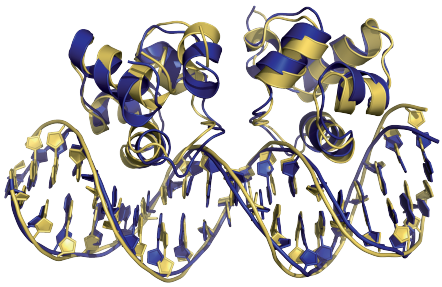



**Figure 5.1.** Illustration of "easy" (interface RMSD < 2.0Å), "intermediate" (2.0Å ≤ interface RMSD < 5.0Å) and "difficult" (interface RMSD ≥ 5.0Å) test cases from the protein-DNA benchmark. "Easy" test case: the Papillomavirus replication initiation domain E-1 (PDB id 1ksy) (interface RMSD = 1.6Å) (**A**). "Intermediate" test case: the intron-encoded homing endonuclease I-PpoI complex (PDB id 1a73) (interface RMSD = 4.3Å) (**B**). "Difficult" test cases: the proline utilization transcription activator (PDB id 1zme) (interface RMSD = 5.8Å) (**C**) and the PVUII endonuclease complex (PDB id 1eyu) (interface RMSD = 6.8Å) (**D**). The bound form of the complex is shown in yellow and the unbound protein in blue. The bound- and canonical B-form DNA structures are shown as insets to highlight the conformational changes in the DNA.

**easy**

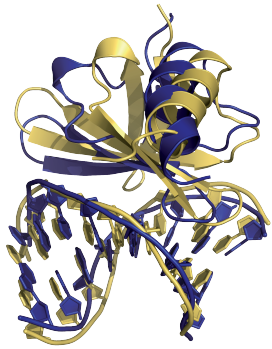


1by4** 0.40[a] 3.55[b] 1.50[c]

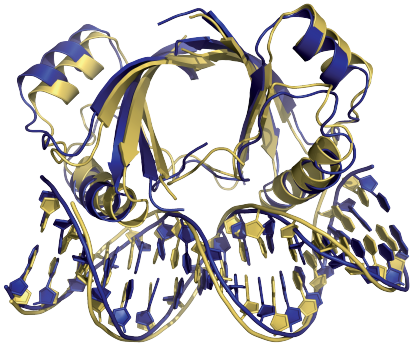


3cro** 0.50[a] 2.23[b] 1.93[c]

**intermediate**

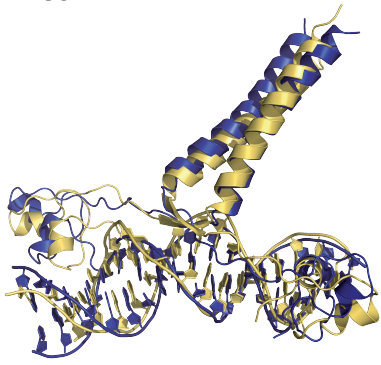


1azp* 0.11[a] 3.44[b] 1.58[c]

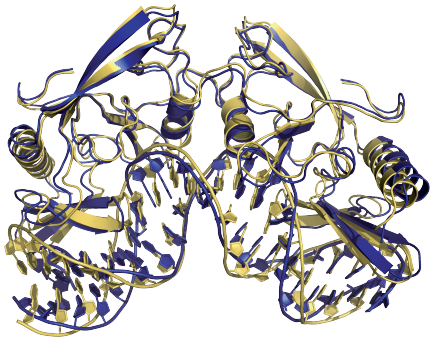


1jj4** 0.44[a] 2.63[b] 2.26[c]

**difficult**



1zme* 0.15[a] 3.75[b] 3.23[c]



1a74** 0.31[a] 3.24[b] 3.70[c]

**Figure 6.5.** Best solutions from unbound flexible docking using an ensemble of custom-built DNA structural models (blue) superimposed on to the reference structure (yellow). The complexes are grouped according to their docking difficulty ("easy", "intermediate" and "difficult) as indicated in the benchmark. De CAPRI score for each solution is indicated as one or two stars after the PDB code as well as the fraction of native contacts (a), the interface (b) and DNA RMSD (c) from the reference structure. RMSD values (Å) were calculated after superimposition on all heavy atoms of the selected regions. The figures were generated using Pymol (DeLano Scientific LLC, www.pymol.org).

DNA interacting proteins are of vital importance to the cell, they regulate the flow of genetic information and guard the DNA against damage. Disruption of these interactions may lead to severe developmental diseases and cancer. An understanding of protein-DNA interactions at atomic detail is, therefore, important.

Computational docking has emerged as a powerful method to model the interactions between proteins and DNA and contribute to the study of these complexes. However, method development is slowed down by the difficulty of locating the interaction interface and the intrinsic flexibility of the DNA helix.

This thesis describes a two-stage protein-DNA docking approach using the HADDOCK docking software. This docking program uses experimental information to drive the docking, facilitating the reconstruction of the correct interaction interface. The combination of explicit flexibility in HADDOCK with implicit flexibility by means of a DNA modelling stage, makes this approach better suited to deal with large DNA conformational changes.