

Последнее обновление: 09/09/2002 04:24:36

# Protein-DNA complexes

## An overview of the structures of protein-DNA complexes

Nicholas M Luscombe, Susan E Austin, Helen M Berman, Janet M Thornton  
Genome Biology 2000 1(1): reviews001.1-001.37

<http://genomebiology.com/2000/1/1/reviews/001>

### Constructing the classification

**ДНК-связывающие белки** играют центральную роль во всех аспектах геномической активности в организме. В последнее время достигнуты определенные успехи в определении структуры ДНК-связывающих белков.

Классификация белок-ДНК комплексов базируется на структуре ДНК-связывающих областей в белках. Таксономия впервые предложена Higgins [1] и модифицирована Ivaldi [2]. Здесь предлагается оригинальная классификация, позволяющая включать новые структуры. Проверка генов, функционально представленных в REBANT [3] показывает, что обычно 2-3% прокармитического генома кодируют ДНК-связывающие белки. Следовательно, предложенная классификация полезна.

**Табл.1** The groups of protein structures found in the dataset, the number of families within each group and the number of PDB files each family contains.

**Табл.2** List of the 240 structures of protein-DNA complexes in the dataset



**Рис.1** | Group I, HTH proteins



**Рис.2** | Group II, zinc-coordinating proteins



**Рис.3** | Group III, zipper-type proteins



**Рис.4** | Group IV, other  $\alpha$  helix proteins



**Рис.5** | Group V,  $\beta$ -sheet proteins

Белок-ДНК комплексы распознавались с помощью X-ray кристаллографии с разрешением более чем в 3.0 Å Protein Data Bank (PDB) [4,5] и Nucleic Acid Database (NDB) [6]. Комплексы устанавливались как структура, содержащая одну или более белковых цепей И, по крайней мере, двунитчатую ДНК длиной более 4-х пар оснований(bp). Исключены однонитчатые и четырехнитчатые ДНК и non-contiguous ДНК (с разрывами нити). В результате выявлено 240 белок-ДНК комплексов (Табл 1, 2). Вох 1 показывает процесс селекции.

Включены 24 гомодимерных комплекса, что асимметричные единицы поддержат только половину структуры.

### Structural taxonomy and classification of protein-DNA complexes

База PDB классифицирована в соответствии со структурой белка в комплексе. Классификация двух-уровневая. На первом уровне белки были рассортированы вручную на 8 групп, используя RasMol [7] данные литературы. Члены одной и той же группы имели характерные общие структурные свойства, используемые для распознавания ДНК, и были родственны друг другу в разной степени. 8 групп были следующими: (I) HTH (including 'winged' HTH), (II) zinc-coordinating, (III) zipper-type, (IV) other  $\alpha$  helix, (V)  $\beta$  sheet, (VI)  $\beta$  hairpin/ribbon, (VII) другие и (VIII) энзимы (Табл. 2). Группа энзимов являлась исключением для структурного критерия, т.к. она содержала все белки, которые обладают ферментативной активностью, если связаны с ДНК. 5 энзимов кроме того классифицированы по структурным основаниям для HTH и др.  $\alpha$  helix групп: restriction endonuclease *FokI* (PDB entry 1fok),  $\gamma$ -resolvase (1gdt), Hin recombinase (1hcr), Tc3 transposase (1tc3) and Cre recombinase (1crx). Эти белки представлены в группе HTH в Табл. 2 и соотв. маркированы.

На втором уровне классификации ДНК-распознающие домены были классифицированы в гомологичные семейства путем сравнения их структуры в парах, используя программу сравнения вторичной структуры SSAP [8]. Программа использует метод динамического программирования [9] и выявляет сходство между белками путем сравнения структурных условий составляющих (constituent) аминокислот. SSAP дает показатель 100 для идентичных белков и более 80 для гомологичных белков; белки автоматически попадают в одно и то же семейство, если показатель выше этой границы (cut-off). Более удаленные белки, которые имеют показатель более 70 также помещаются в одно и то же семейство, если они выполняют сходные биологические функции [10].

Белки разрывались по их составляющим ДНК-связывающим доменам, прежде чем проводилось сравнение. В большинстве димеров каждый домен, соответствует определенной субъединице и структуре, которую необходимо отделить в constituent цепь. У белков, таких как те, что с  $\beta\beta\alpha$  zinc fingers, однако, цепь содержит несколько связывающих доменов; в каждом случае, следовательно, субъединицы подразделяли на соответствующие сегменты, которые представлены в Табл. 2. Структуры идентифицировались стандартным four-digit PDB кодом (напр., 1aay). Когда белковая субъединица специфицировалась, то соответствующая идентификация цепи в PDB файле добавлялась к four-digit коду (напр., 1aayA). Для определенного сегмента внутри субъединицы добавлялся числовой идентификатор, как показано в Табл. 2, (напр., 1aayA1).

Всего представлено 54 семейств белков, из которых 33 имеют более одного PDB entry. Внутри каждого семейства имеется структура того же самого белка, связанного с различными последовательностями ДНК (напр., phage 434 repressor complexes 1per and 1rpe в Cro и Repressor семействе) и структуры



(Рис.6.) | Group VI, the  $\beta$ -hairpin/ribbon proteins



(Рис.7.) | Group VII, other ДНК-binding proteins



(Рис.8.) | Group VIII, the enzymes



(Box.1.) | Flow diagram showing the selection of the protein-ДНК complexes from the PDB (04/01/00).

разных белков, связанные с разными последовательностями ДНК, (напр., phage 434 и  $\lambda$  repressor complexes, 1per и 1lli соответственно, в Cro и Repressor семействе). Табл. 2 представляет список всех структур белок-ДНК комплексов и их классификацию. Показаны также множественные сравнения последовательностей ДНК, которые связываются в каждом семействе ( ClustalX [11]).

## Group I: helix-turn-helix proteins

**Мотив НТН** является наиболее распространенным элементом распознавания, используемым транскрипционными регуляторами и ферментами у прокариот и эукариот. Хотя мотив традиционно определяется как сегмент в 20 аминокислот из двух почти перпендикулярных  $\alpha$  спиралей, соединенных с помощью четырех остатков  $\beta$  см (Cro и Repressor family, 1lmb; Рис. 1a) авт. расширили это определение для тех с более длинными линкерами, такими как петли, такой же длины относительно ориентации  $\alpha$  спиралей (напр., белок семейства RAP1, 1ign). Примеры для каждого семейства в группе НТН показаны на Рис. 1. На рис. НТН выделена красным.

**Мотив** всегда связывается с большой бороздой ДНК; вторая  $\alpha$  спираль, обычно известная как распознающая или зонд, вставляется в борозду. В большинстве комплексов прямые контакты совершаются между аминокислотными боковыми цепочками и нуклеотидными основаниями; в немногих случаях, однако, используются атомы белкового остова или мостики из молекул воды (напр., Trp семейство репрессоров, 1trrA). Подтвержденные контакты с ДНК остовом в основном формируются линкером и первой  $\alpha$  спиралью в мотиве, который соединяет мостиком большую борозду на N-терминальном конце распознающей спирали. Дальнейшие взаимодействия с нуклеиновой кислотой м. также совершаться с помощью остальной части белка и иногда вносят дальнейшую спецификацию ДНК последовательностей. Напр., Hin recombinase белок (1hcrA) взаимодействует с основаниями в большой борозде рядом с одной из связей, образуемых спиралью распознавания.

**Мотив НТН** обычно обнаруживается в пучке из 3-6  $\alpha$  спиралей, которые создают стабилизирующую гидрофобную сердцевину. Хотя мотивы из разных семейств белков структурно очень сходны незначительная гомология обнаруживается вне мотива. В структурах, таких как 434 репрессорный белок (1lli, семейство Cro и Repressor), НТН мотив является частью основного тела белка. У др., таких как пуриновый репрессор, (1wet, семейство LacI репрессоров), он находится в небольшом домене, отходящем от основной структуры. Имеется незначительное сходство последовательностей между мотивами разных семейств и эта изменчивость позволяет им распознавать определенные наборы последовательностей ДНК.

**Точное** расположение распознающей спирали в большой борозде ДНК также варьирует, отражая структурную и функциональную потребности каждого белка. Распознающая спираль прокариотических транскрипционных факторов (напр., из семейств Cro и Repressor, таких как 1lli) обычно расположена параллельно base-pairing краям нуклеотидов, тогда как эта спираль у эукариотических белков (напр., гомеодоменное семейство, см 1oct) расположена параллельно сахар-фосфатному остову, чтобы приспособиться к более длинным  $\alpha$  спиральям. связывание с помощью Trp репрессора (1trrA) уникально, т.к. N-терминальный конец  $\alpha$  спирали практически оказывается в борозде. Хотя такое расположение ограничивает роль аминокислот, расположенных ниже  $\alpha$  спирали, оно необходимо, чтобы допустить вторую репрессорную субъединицу в ту же самую большую борозду, если связывание в тандеме. Связывание спирали в большой борозде, которая также очень распространена и в др. группах, создает геометрически благоприятные условия, в которых компоненты как белка, так и ДНК м. меняться, делая возможной мультиспецифическую комплементарность. Белковые последовательности и способы взаимодействия варьируют существенно, но нуждаются в спирали, чтобы быть 'представленными' на поверхности белка, готовыми для взаимодействия с ДНК, это и обеспечивается НТН мотивом.

**В** целом прокариотические транскрипционные факторы связывают палиндромные последовательности ДНК как гомодимеры, тогда как эукариотические белки, такие как члены гомеобоксного семейства, связывают как мономеры, так и гетеродимеры с несимметричными сайтами-мишенями.

Последнее расположение потенциально позволяет распознавать значительно более широкий круг последовательностей ДНК. Прокариотические энзимы в группе (напр., *FokI* endonuclease, *1fok*), которые функционируют как мономеры, обладают более, чем одним мотивом в одиночно субъединице.

Имеется 16 гомологичных семейств в группе НТН. 8 содержат только по одной структуре каждый, а из оставшихся 6 только Cro и Repressor и homeodomain семейства содержат белки с разными аминокислотными последовательностями. Идентичность парных последовательностей между субъединицами в семействах Cro и Repressor колеблется в пределах от 68% (1lmbA и 1regA) до 100% для идентичных белков (1lilA и 1lmbA). Попарные SSAP показатели выше 85. В гомеодоменовом семействе, хотя POU доменовые белки часто рассматриваются отдельно, они были объединены вместе в данной работе из-за их высокого показателя SSAP. Напр., белок Mata-2 (1aplA) и POU доменовый белок Pit-1 (1au7A1) имеют показатель SSAP 88.3 при сравнении 59% белковых остатков. Как результат, более значительная изменчивость при попарной идентификации последовательностей, которая самой низкой была в 42% (1aplA и 1au7A1). *Hin* recombinase,  $\gamma\delta$ -resolvase, *FokI* restriction endonuclease, Tc3 transposase и Cre recombinase семейства принадлежат как НТН так и энзимной группе.

### 'Winged' НТН proteins

Мотив 'winged' НТН является расширением группы НТН, который характеризуется присутствием третьей  $\alpha$  спирали и соседнего  $\beta$  листка (Рис. 1m-p), которые рассматриваются как компоненты ДНК-связывающего мотива. Распознающая спираль связывается как в регулярных НТН мотивах, так и в добавочных вторичных структурных элементах, обеспечивая дополнительные контакты с остовом ДНК.

### Group II: zinc-coordinating proteins

Zinc-coordinating белки образуют одну большую группу транскрипционных факторов в геноме эукариот, а ДНК-связывающий мотив характеризуется четырехгранной координацией одного или двух ионов цинка за счет законсервированных цистеинового или гистидинового остатков. Широко распространенное использование этой организации, как полагают, обусловлено структурной стабильностью ионов металла, принадлежащих доменам, которые недостаточно велики для образования стабильного гидрофобного ядра. Использование zinc-coordinating мотивов не ограничено связыванием ДНК и они обнаруживаются также в доменах, которые обеспечивают межбелковые взаимодействия. Белки в этой группе структурно более различны, чем белки НТН группы и идентифицировано 6 принципиальных семейств, из которых 4 представлены в базе данных комплексов. Характерные структуры представлены на Рис. 2, где zinc-coordinating мотив окрашен красным. Чтобы избежать недопонимания при использовании термина "цинковые пальчики" ('zinc finger'), мы зарезервировали его использование для белков, которые имеют Zif-268-style (1aayA1) мотив с двумя  $\beta$  нитями и с  $\alpha$  спиралью (Рис. 2a). Термин 'zinc-coordinating' будет использоваться как обобщающий термин для всех белков с ионами цинка в ДНК-связывающем мотиве.

### The $\beta\beta\alpha$ zinc-finger family

$\beta\beta\alpha$  zinc-finger белки составляют самое большое индивидуальное семейство в группе и более тысячи мотивов с определенными последовательностями идентифицированы в транскрипционных факторах. Структура пальчиков характеризуется коротким двунитчатым антипараллельным  $\beta$  листком, сопровождаемым  $\alpha$  спиралью (Рис. 2a). Две пары законсервированных гистидиновых и цистеиновых остатков в  $\alpha$  спирали и второй  $\beta$  нити скоординированы с единственным ионом цинка.

Белковые субъединицы часто содержат множественные пальчики, которые оборачиваются вокруг ДНК в виде спирали. Пальчики связывают соседние в 3 п.н. субсайты путем проникновения  $\alpha$  спирали в большую борозду и благодаря

паттерну распознавания между спиралью и ДНК. Аминокислоты в положении -1, 2, 3 и 6 относительно точки старта спирали  $\alpha$  используются для взаимодействия с основаниями, находящимися в положении -1, которое предшествует спирали. Хотя имеются примеры комплексов, которые не следуют этому порядку, было установлено, что повреждение аминокислот в ключевой позиции ведет к тому, что начинают распознаваться последовательности др. субсайта. При образовании нескольких пальчиков в белке связывающие сайты различной длины м. связываться с различной специфичностью. Напр., белок с 5 пальчиками м. ожидать, будет связываться с длинным сайтом-мишенью очень избирательно, тогда как белок только с одним пальчиком будет потенциально связываться с широким кругом сайтов, содержащим необходимые последовательности в субсайте. Однако, структура белка глиобластомы человека (1gli) указывает на то, что связывание не всегда столь прямолинейно; из 5 пальчиков в структуре один не контактирует с ДНК вовсе, и лишь два, по-видимому, образуют специфические контакты с основаниями.

Как уже указывалось, белковые субъединицы подразделены на определенные домены, каждый содержит мотив цинковых пальчиков. Парная идентификация последовательностей сравниваемых доменов обнаруживает высокое сходство от 73% (напр., для белка zinc-finger человека, 1udbA1, и белка *Drosophila* tramtrack, 2drpA1) до 100% (напр., мышинный Zif268 белок, 1aayA1, и искусственный белок, 1meu). Все домены структурно очень сходны, показатель их SSAP выше 90.

### Hormone receptor family

Ядерные рецепторы для стероидных гормонов, тироидных гормонов и ретиноидов образуют второе семейство в группе (Рис.2b). В ответ на связывание соответствующего лиганда эти рецепторы транслоцируются из цитоплазмы в ядро и регулируют транскрипцию последовательностей ДНК, называемых hormone response elements. Гормональные рецепторы обычно функционируют как гомо- или гетеро-димеры, а каждый мономер обычно состоит из лиганд-связывающего, ДНК-связывающего и регулирующего транскрипцию доменов. Мотив zinc-coordinating обнаруживается в ДНК-связывающем домене и характеризуется двумя антипараллельными  $\alpha$  спиралями, покрытых петлями их N-терминальных концов; каждая пара спираль-петля координирует один ион цинка, используя 4 законсервированных цистеина. Две  $\alpha$  спирали расположены примерно под прямым углом др. к др.; первая проникает в большую борозду ДНК, чтобы осуществить взаимодействия с основаниями, тогда как петля и вторая  $\alpha$  спираль контактирует с остовом ДНК. Одного ДНК-связывающего домена достаточно для димеризации, а сторона (interface), образуемая петлей, ведет во вторую  $\alpha$  спираль.

Все рецепторные субъединицы распознают одну из двух последовательностей полу-сайта, 5'-AGAACA-3' или 5'-AGGTCA-3'. Идентификация полного сайта-мишени предопределяется двумя последовательностями полу-сайтов, которые присутствуют, их относительной ориентацией (или симметричной или палиндромной) и расстоянием между ними (3-6 п.н.). Т.обр., распознавание последовательностей мишени зависит от читабельности последовательностей полу-сайтов каждой субъединицей и от способа, с помощью которого субъединицы димеризуются. Последовательности всех данных в имеющейся базе данных очень сходны (их идентичность > 90%). за исключением рецепторов тироидного гормона (напр., 1bsx), который имеет две дополнительные спирали в С-терминальном хвосте. Структуры всех очень сходны, попарный показатель SSAP свыше 90.

### Loop-sheet-helix family

Третье семейство с мотивом zinc-coordinating является семейство с мотивом loop-sheet-helix zinc-coordinating (Рис. 2c). Оно представлено ДНК-связывающей областью белка p53, активатора транскрипции, участвующего в супрессии опухолей. Его ДНК-связывающий домен состоит из петли, отходящей от основного тела белка, сопровождаемой небольшим  $\beta$  листком, из  $\alpha$  спирали и затем еще одной петли, которая идет обратно в белок. Три цистеина и гистидин в двух петлях координируют ион цинка.

Белок связывается  $\alpha$  спиралью с большой бороздой ДНК, а петля с

минорной бороздой, хотя последняя вряд ли обладает специфичностью. Белок функционирует как тетрамер, каждая его субъединица контактирует с отдельными распознаваемыми последовательностями в 5 п.н., расположенными одна за другой. Области вне ДНК-связывающего мотива образуют взаимодействия между субъединицами.

### Gal4 family

**И**, наконец, семейство, состоящее только из одного Gal4 белка. Это транскрипционный регулятор генов, индуцируемых галактозой, а его zinc-coordinating мотив идентифицирован еще только в белках дрожжей. Мотив этот представлен парой  $\alpha$  спиралей, которые координируют ионы цинка посредством 6 цистеиновых остатков, где два цистеина связаны с двумя атомами металла (Рис. 2d). Первая  $\alpha$  спираль проникает в большую борозду ДНК для связывания с основаниями, а вторая  $\alpha$  спираль взаимодействует с остовом. Gal4 функционирует как гомодимер, а стороны для димеризации расположены вне zinc-coordinating мотива.

### Group III: zipper-type proteins

Группа zipper-type получила свое название по способу димеризации ее членов, она характерна только для эукариотических организмов. Известны два семейства лейциновых застежек (leucine zipper) (Рис. 3a) и helix-loop-helix белков (Рис. 3b); последнее нельзя путать с НТН группой. Т.к. некоторые члены функционируют в качестве гетеродимеров (напр., Fos-Jun complex), все данные PDB представлены только гомодимерами.

### Leucine zipper family

**В** семействе лейциновых застежек структура белка м.б. подразделена на две части: область димеризации и ДНК-связывающая область. Как показано на Рис. 3a, каждая субъединица в leucine zipper белках состоит из одной петли примерно в 60 аминокислот. Димеризация обеспечивается за счет образования суперскрученной секции в 30 аминокислот на С-терминальном конце каждой спирали. Сегмент, известный как область застежки, состоит из лейцина или сходных гидрофобных аминокислот в каждой восьмой позиции - примерно каждые два витка  $\alpha$  спирали. Соответствующие боковые цепочки от каждой субъединицы обеспечивают гидрофобные контакты сторон посредством бок-о-бок упаковки. ДНК-связывающая область, известная также как basic область, находится на N-конце, а связывающий сегмент управляет расширением области димеризации.  $\alpha$  спирали двух субъединиц дивергируют из суперскрученного участка и проникают в большую борозду ДНК в противоположных направлениях, каждая связывает половину мишени. Семейство leucine zipper представлено целиком дрожжевыми GCN4 белками, которые имеют почти идентичную структуру и связываются с промоторными областями генов, кодирующих энзимы, участвующие в биосинтезе аминокислот.

**Белки helix-loop-helix** являются модификацией непрерывных  $\alpha$  спиралей белков leucine zipper, в которых ДНК-связывающая и димеризации области разделены петлей, дающих в результате четырехспиральный пучок (Рис. 3b). Подобно лейциновым застежкам спирали димеризации взаимодействуют др. с др. в суперскрученном образовании, а ДНК-связывающие спирали вставляются в большую борозду ДНК. Разделение двух сегментов создает большую гибкость и позволяет позиционировать спирали-зонды на нуклеиновых кислотах. Семейство helix-loop-helix представлено мышинными и человеческими формами Max, мышинными MyoD и человеческими USF белками. Идентичность последовательностей колеблется в пределах от 66% (Max белок, 1a2A, и USF белок, 1a4A) до 97% (мышинный Max белок, 1a2A, и человеческий Max белок, 1h0A) и за исключением MyoD (1mdyA) и USF (1a4A) белковой пары (парный SSAP показатель 70), SSAP показатели выше 80. Структурные различия между белками в основном возникают из-за изменчивости длин и расположения петель.

### Group IV: other $\alpha$ -helix proteins

Известно 7 семейств с очень различными функциями в 'other  $\alpha$  helix' группе (Рис. 4). Skn-1 (1skn; Рис. 4d) и MADS (Рис. 4g для MADS box, 1mnm) являются регуляторными областями транскрипции в эукариотических белках, papillomavirus-1 E2 (2bor; Рис. 4a) и EBNA1 (1b3t; Рис. 4c) являются регуляторами транскрипции и инициаторами репликации, гистоны (1aoi; Рис. 4b) и high-mobility group (HMG) белки (1qrv; Рис. 4f) являются архитектурными белками для упаковки ДНК и Cre (1crx; Рис. 4e) является сайт-специфической рекомбиназой. Хотя структуры белков очень отличны, все они используют  $\alpha$  спирали (окрашены красным на Рис. 4) как основной способ связи с ДНК.

Skn-1 и MADS белки связывают длинные зонд(probe)-спирали в большой борозде ДНК способом, сходным с zipper-type белками. Skn-1 (Рис. 4d) является мономером с компактной четырех-спиральной единицей; самая длинная  $\alpha$  спираль на С-терминальном конце связывается с большой бороздой, а остальная часть домена контактирует с основанием ДНК. У MADS (Рис. 4g), антипараллельный  $\beta$  листок и соседний суперскрученный участок образуют поверхность димеризации.  $\alpha$  спираль на противоположной стороне листка дивергирует от центра связывающего сайта в соседнюю большую борозду, к контактирующим основаниям и группам остова. ДНК изгибается по направлению белка.

Papillomavirus-1 E2 и EBNA1 являются структурно сходными димерными белками, которые м.б. подразделены на две области (Рис. 4a, 4c). В стержневой области, четыре  $\beta$  нити от каждой субъединицы комбинируются в 8-нитчатый  $\beta$  цилиндр (barrel). Фланги ДНК-связывающих областей проецируют одиночную спираль  $\alpha$  симметрично в большую борозду ДНК. Исходя из их структуры (Рис. 4a, 4c), связывающие ориентации спиралей очень различны в двух семействах.

Histone и HMG являются мультимерными белками, которые связывают ДНК независимо от последовательностей оснований. Гистон (Рис. 4b) октамерный белок, чья структура м.б. почти цилиндрической. Каждая субъединица представлена пучком из трех или четырех спиралей, которые упакованы одна напротив другой; длинный сегмент ДНК обвивается вокруг циркулярного края белка. Соседняя  $\alpha$  спираль образует экстенсивные контакты с группами остова ДНК чтобы стабилизировать искажения, но ни одна не проникает в борозду и происходит, следовательно, лишь немного взаимодействия с основаниями. Субъединица HMG представлена тремя  $\alpha$  спиральями, которые расположены в виде L (Рис. 4f). первая и вторая спирали связывают основания и группы остова из минорной борозды и вызывают серьезные искажения структуры ДНК благодаря интеркаляции аминокислотных боковых цепочек.

Наконец, Cre (Рис. 4e) является димерным белком. Каждая субъединица состоит из двух структурных доменов, которые складываются в сложный пучок спиралей. Совместно домены формируют зажим вокруг ДНК, а спирали проникают в большую и малую борозды.

### Group V: the $\beta$ -sheet proteins

Группы V и VI представлены белками, которые используют  $\beta$ -нитевые структуры для распознавания и связывания ДНК. Группа V, которая только одна содержит TATA box-связывающее семейство белков, характеризуется использованием широкого  $\beta$  листка для связывания ДНК (Рис. 5).

TATA box-связывающие белки являются существенным компонентом мультисубъединичных комплексов, иницирующих транскрипцию, который собирается на промоторах генов, которые транскрибируются с помощью РНК полимеразы II. Хотя они являются молекулами из одиночной цепи их структуры рассматриваются обычно как состоящие из двух псевдо-идентичных доменов. 10-нитчатый антипараллельный  $\beta$  листок, соединяющий домены, покрывает минорную борозду ДНК; он образует два two надежных перегиба от основного тела белка путем интеркаляции фенилаланиновых боковых цепочек от любого конца листка. Семейство представлено белками от бактерий *Pyrococcus woesei*, дрожжей и человека. Сравнение последовательностей и структур различных субъединиц дают очень высокий показатель SSAP (>90% и >90 соответственно).

## Group VI: the $\beta$ -hairpin/ribbon proteins

Члены этой группы отличаются от TATA box-связывающих белков, они используют очень маленькие двух и трех-нитчатые  $\beta$  листки или мотивы шпилек для связи или с большой или малой бороздой ДНК (Рис. 6). 6 семейств белков с очень разными функциями: MetJ repressor (1cma; Рис. 6a), Arc repressor (1bdt; Рис. 6f) и T-доменное семейство (1xbr; Рис. 6d) транскрипционных регуляторов; integration host factor (1ihf; Рис. 6c) и hyperthermophile хромосомные белки (1azp; Рис. 6e), действующие как поддерживающие леса, чтобы диктовать структуру ДНК для образования комплексов белок-ДНК высокого порядка; Tus белок (1ecr; Рис. 6b) завершающий репликацию ДНК с помощью helicases. Хотя в целом структуры белков отличны, имеются общие темы в использовании  $\beta$  нитей.

MetJ и Arc репрессоры являются димерами с очень сходным способом связывания (Рис. 6a, 6f). Каждая белковая субъединица представлена пучком спиралей и одиночной  $\beta$  нитью; нити от каждой субъединицы упаковываются сторона к стороне, формируя антипараллельный листок, который связывается с большой бороздой ДНК. Листок лежит плоско в борозде; следовательно, белковые боковые цепи от непосредственно одной лицевой стороны нити взаимодействуют с краями основания.

Терминатор репликации Tus и T-доменные белки используют  $\beta$ -нитевые мотивы, чтобы связаться с большой бороздой ДНК (Рис. 6b, 6d). В обоих, нити расположены почти перпендикулярно краю основания, что делает возможным контакты аминокислот, которые экспонируют свои боковые цепи на любую лицевую сторону листка. Терминатор репликации Tus является мономерным белком образующим N- и C-терминальные  $\alpha$ -спиральные пучки, которые соединены с помощью антипараллельных bundles  $\beta$  нитей. Структура образует большую щель, в которой ДНК связывается с нитями, обращенными к большой борозде. Напротив, T-домен связывается как димер. Каждая субъединица состоит из  $\beta$  цилиндра: один конец цилиндра нацелен в направлении ДНК и представлен двумя  $\beta$  нитями, одна из которых распространяется в большую борозду.

Как integration host factor так и хромосомные белки связываются с минорной бороздой и искажают ДНК путем интеркаляции боковых цепей от мотивов  $\beta$  листка (Рис. 6c, 6e). Integration host factor действует как димер; плечо  $\beta$ -шпильки от каждой субъединицы распространяется в направлении оппозитной стороны ДНК и вставляет пролиновые боковые цепи между определенными base-steps. Минорная борозда расширяется в области связывания и ДНК наклоняется в направлении основного тела белка. Напротив, hyperthermophile хромосомный белок действует как мономер и использует трехнитчатый  $\beta$  листок для связи с минорной бороздой. Две гидрофобные боковые цепочки от соседних нитей интеркалируют в одиночную base-step, вызывая отклонение ДНК от белка.

Только семейства хромосомных белков и Arc репрессора содержат более одной структуры. Парная идентификация последовательностей и SSAP показатель между субъединицами внутри семейств высоки (>90% и >90% соответственно).

## Group VII: other

Имеется два не энзиматических семейства с современной базе данных, которые не используют хорошо определяемый вторичный структурный мотив для связывания ДНК (Рис. 7). Оба функционируют как димеры и имеют мультидоменные субъединицы, которые обеспечивают ДНК-связывание, димеризацию и локализацию в ядре. Эти белки е proteins покрывают нуклеиновую кислоту, а комплексы симметричны, если рассматривать параллельно длинной оси ДНК. Межнитевые и внутримолекулярные петли обеспечивают большую часть контактов с основаниями и остовом.

Область Rel гомологии (Рис. 7a) является законсервированным N-терминальным доменом транскрипционных регуляторов, участвующих в клеточной защите и дифференцировке. Каждая субъединица представлена двумя  $\beta$ -sandwich доменами, которые соединены с помощью более чем 10 межнитевых петель, которые связываются в большой борозде ДНК. Семейство STAT (Рис. 7b) содержит транскрипционные факторы, которые обеспечивают реакцию на цитокины и ростовые факторы. Каждая белковая субъединица

состоит из четырех структурных доменов, а функциональный димер напоминает пару игроков с ДНК связью в шарнире. Окружающие петли и  $\alpha$  спираль приближают ДНК из большой и малой борозд.

### Group VIII: the enzymes

Энзимная группа завершает классификацию базы данных (Рис. 8). Белки в энзимную группу объединены на основании их функции; все они меняют структуру ДНК благодаря катализу химических процессов.

ДНК-связывающие области, используемые энзимами в целом довольно жестки для описания их в терминах простых структурных мотивов и эти белки используют экстенсивную комбинацию  $\alpha$  спиралей,  $\beta$  нитей и петель для распознавания и связывания ДНК (Рис.8). Большинство энзимов представлено тремя четкими доменами: ДНК-распознающий домен, который 'считывает' последовательности ДНК; каталитический домен с энзиматически активным сайтом; и, там где необходимо, домен димеризации, однако имеются и исключения. Структура часто при этом образует U-образную полость с которой ДНК связана и часто структура ДНК при этом деформируется будучи связанной.

Для сиквенс-специфичных энзимов последовательностями-мишенями обычно являются 4-8 п.н., и связывание является более избирательным, чем у транскрипционных регуляторов. Напр., в белках, таких как *HhaI* methyltransferases и endonucleases, одиночная замена в последовательности-мишени м. вести к редукции связывающей активности более чем в миллион раз. Белки, как полагают, приобретают свою специфичность как от считывания последовательностей оснований, так и каталитического действия на ДНК, как это имеет место у endonucleases *BamHI* (3bam; Рис. 8e) и *EcoRI* (1qps; Рис. 8d), или даже в первую очередь благодаря каталитическим процессам, как это имеет место у endonuclease *EcoRV* (1rva; Рис. 8c). Др. белки, такие как полимеразы должны, однако, обеспечивать сиквенс-независимые взаимодействия с из ДНК субстратом, все еще сохраняя специфичность правильно различать пары оснований от неправильных последовательностей. 7 endonucleases и 4 polymerases (см. семейства 40-46 и 47-50, и Рис. 8b-8h и Рис. 8i-8l, соотв.), доминируют в этой группе из 16 семейств.

### A protein-ДНК complex website

Website, который суммирует группы и семейства protein-ДНК комплексов находится по адресу [http://www.biochem.ucl.ac.uk/bsm/prot\\_dna/prot\\_dna.html](http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna.html). Он включает краткое описание каждого семейства и информацию, установленную для субъединиц каждого белка, структурную информацию, таблицы попарной идентификации последовательностей и показатели SSAP. Приведены связи белков с их соотв. PDB и NDB данными и PRINTS анализ последовательностей мотивов. Кроме того приведены ссылки на PDBsum, нашу базу данных, суммирующую и анализирующую структуру файлы PDB. Каждая структура содержит информацию о ее CATH, PROCHECK и PROMOTIF анализе и ссылки на SCOP, WHATIF check и FSSP структурные данные.

### Conclusions

Представленные данные создают базу для лучшего понимания образования комплексов белок-ДНК. Они проливают свет на разнообразие таких комплексов и подчеркивают важность взаимодействий между  $\alpha$  спиралью и большой бороздой, что является основным способом связи в 28 из 54 семейств. В частности, HTH и zinc-coordinating мотивы используются повторно и создают компактный каркас, который представляет  $\alpha$  спираль на поверхности структурно различающихся белков, готовую для взаимодействия с ДНК. Эти структуры обнаруживают множество вариаций как в аминокислотных последовательностях, так и в деталях геометрии и задействованы в соответствии с потребностями контекста, в котором они находятся. Для достижение тесного соответствия между  $\alpha$  спиралью и большой бороздой имеется достаточно гибкости, чтобы и белок и ДНК адаптировали соответствующие конформации, которая обеспечивает мультиспецифичную комплементарность. Каждое из этих взаимодействий базируется не на простом

коде, связанном с аминокислотными последовательностями и последовательностями ДНК, с которыми они соединяются. Принимая во внимание дополнительную сложность целиком разных каркасов, теперь становится ясным, что детальные правила для распознавания оснований ДНК д.б. специфичными для семейств, но в их основе д. лежать тенденции, такие как arginine-guanine взаимодействия.

**В**ывяляются и различия между белковыми доменами, которые непосредственно связываются с ДНК, и те, что участвуют в катализе. Хотя имеются исключения, но сначала происходит притяжение ДНК одной из сторон и входом в борозду для взаимодействия с краями оснований. Последние обычно составляют субстрат, используемый сложными сетями вторичных структур и петель. часто вызывающих значительные искажения ДНК - обычно необходимые для каталитического процесса. Способность изгибать ДНК, однако, не ограничена только энзимами; хотя и не столь сильное, но четкое изгибание ДНК является также общераспространенным свойством комплексов, образуемых транскрипционными факторами. Этот и др. эффекты, такие как электростатичность, водой- и катионами-опосредованные взаимодействия способствуют косвенно распознаванию последовательностей ДНК.

**В** табл. 1, суммированы данные в основном по структуре эукариотических белков. Они демонстрируют, что ДНК-связывающие домены имеют очень значительное структурное разнообразие, чем др. Это не является неожиданностью, учитывая, что эти организмы сформировали довольно сложные транскрипционные и репарационные механизмы, и кроме того, большинство эукариотических белков выявлены и структурно охарактеризованы. Безусловно должны существовать и др. еще не описанные способы связи белков с ДНК. Геномный анализ не только позволит идентифицировать такие белки, но и позволит определить функционально важные сайты-мишени на ДНК. Outline of the families of ДНК-binding proteins

### A complete outline of the families of ДНК-binding proteins and their functional, structural and binding properties follows.

**Box 1** shows the selection process by which the dataset was compiled. Table 1 provides a summary of the families and Table 2 lists the 240 structures of protein-ДНК complexes in the database. Figures 1-8 show ribbon diagrams of the relevant structures.

**Group I:** Helix-turn-helix (HTH) group

#### 1. Cro and Repressor family

**Function.** The Cro and Repressor proteins (Figure 1a) are part of the lysogenic/lytic growth switch mechanism in bacteriophages and function as transcriptional regulators at a set of six related operons.

**Structure.** Both protein types function as homodimers. Each Repressor subunit has two domains: an amino-terminal five-helix bundle whose second and third  $\alpha$  helices comprise a HTH motif; and a carboxy-terminal domain that mediates dimerization (Figure 1a). Cro is a single-domain protein with a structure homologous to the amino-terminal region of Repressor. The fourth and fifth  $\alpha$  helices mediate dimerization.

**Binding.** Cro and Repressor bind six related operons with varying affinities. Each operon is 14 bp long and pseudosymmetrical; four bases at either end are conserved between sites and the variation in the sequence of the central 6 bp are thought to modulate the binding affinity of the protein. The recognition helix of the HTH motif contacts base edges in the ДНК major groove.

#### 2. Homeodomain family

**Function.** These are transcription regulators for a wide range of genes; in particular many have a vital role in development and cell differentiation (for example, Mat  $\alpha$ -2; 1apl). Some are expressed broadly whereas others are tissue specific.

**Structure.** The proteins are small (just over 100 amino-acid residues in length) and consist of four helices.

**Binding.** The protein binds ДНК either as a monomer or a dimer, depending on the protein and many are capable of both. Typical HTH binding is displayed in Figure 1b, with the

second helix of the motif inserted in the  $\Delta$ HK major groove.

### 3. LacI repressor family

**Function.** Lac repressor regulates the lac operon, which codes for proteins required to transport and degrade lactose. The purine repressor proteins of the LacI repressor family regulate *de novo* purine and pyrimidine synthesis by repression of genes encoding enzymes that participate in the synthesis pathway. Guanine and hypoxanthine act as co-repressors on binding to the protein. Other members of the LacI repressor family, not represented in the current dataset, display high structural and sequence similarity and control a wide range of biosynthetic pathways.

**Structure.** Purine repressors function as homodimers, as do most other family members (Figure 1c). The lactose, fructose and raffinose repressors are exceptions, and appear to exist as tetramers [67]. Each subunit is a two-domain structure. The amino-terminal domain (approximately 60 residues) contains a three-helix bundle followed by a loop and an additional helix. The first two  $\alpha$  helices form the HTH motif and the fourth is called the hinge helix. The larger carboxy-terminal domain (about 280 residues) is a mixture of  $\alpha$  helices and  $\beta$  strands and binds the co-repressor.

**Binding.** Binding sites are typically 16-18 bp long and pseudo-palindromic. The recognition helix of the HTH motif binds in the major groove and phosphate backbone contacts are mediated by the remainder of the helical bundle. The hinge helix from each subunit is inserted in the same  $\Delta$ HK minor groove at the center of the binding site and jointly introduce a kink by intercalation of leucine sidechains.

### 4. Endonuclease *FokI* family

**Function.** Endonuclease *FokI* is a bipartite restriction enzyme which recognizes a specific  $\Delta$ HK sequence and non-specifically cleaves at a position a short distance away.

**Structure.** The protein acts as a monomer with two functional regions (Figure 1d). The amino-terminal  $\Delta$ HK-recognition region (about 390 residues) may be divided into three further subregions. D1, a roughly 160 residue subregion made of an amino-terminal arm, ten  $\alpha$  helices and a two-stranded  $\beta$  sheet. Helices 5, 6 and 8 form a pseudo-HTH motif. Helices 5 and 6 lie on the same helical axis, jointly forming the first  $\alpha$  helix, and helix 8 acts as the recognition helix. A subregion (D2), of about 110 residues, contains six  $\alpha$  helices and a three-stranded  $\beta$  sheet with the  $\alpha$  helices packing in a triangular formation and the second and fifth  $\alpha$  helices arranged in a HTH-like manner. The turn is replaced by an extensive loop region - D3 - an approximately 80-residue segment containing five  $\alpha$  helices and a three-stranded  $\beta$  sheet. The carboxy-terminal catalytic domain (about 180 residues) is made of a five-stranded  $\beta$  sheet flanked by seven  $\alpha$  helices. The active site is situated on the first three  $\beta$  strands in the region .

**Binding.** Binding is to a site containing the sequence 5'-GGATG-3' and staggered cleavage occurs 9 and 13 bp away from the target sequence. All base contacts to the recognition sequence are made by subregions D1 and D2. The amino-terminal arm and second  $\alpha$  helix from D1 bind in the major groove and a loop preceding this recognition helix is found in the minor groove. The recognition helix from the HTH motif in D2 contacts the major groove. The catalytic region is positioned adjacent to the  $\Delta$ HK-recognition region.

### 5. $\gamma\delta$ -resolvase family

**Function.** The  $\gamma\delta$ -resolvase is a site-specific recombinase which converts negatively supercoiled circular  $\Delta$ HK containing two directly repeated copies of the recombination site into two interlinked rings.

**Structure.** The protein functions as a homodimer (Figure 1e). Each subunit is made of two domains. The amino-terminal domain (about 120 residues) contains the catalytic center and the dimerization interface. It consists of a five-stranded  $\beta$  sheet flanked by three  $\alpha$  helices on one side and a single  $\alpha$  helix on the other. The longest  $\alpha$  helix packs with its counterpart in the other subunit to stabilize the dimer. The carboxy-terminal domain (approximately 40 residues) is a three-helix bundle with the second and third  $\alpha$  helices forming a HTH motif. An extended arm region (about 20 residues), comprising the carboxy-terminal half of the dimerization helix and a loop, connect the two domains.

**Binding.** Each 114 bp recombination region consists of three resolvase-binding sites, I, II and III. Each site binds a resolvase dimer and is made of an inverted repeat of a 12 bp recognition sequence with varying base sequence and spacing between the half-sites. The

structure found in 1gdt (Figure 1e) is thought to represent the conformation found prior to the recombination process. Two main ДНК-binding regions are found in each subunit. The recognition helix of the carboxy-terminal HTH motif binds in the major groove at the outer ends of the binding site. The extended helix in the arm region is inserted in the minor groove near the center of the binding site in a similar manner to the recognition helices from leucine-zipper structures. The ДНК is bent 60° away from the main body of the protein. The ДНК is slightly kinked at the center of the site owing to partial intercalation of threonine residues from the arm region.

#### 6. Hin recombinase family

**Function.** The Hin recombinase protein catalyzes site-specific recombination in the *Salmonella* chromosome.

**Structure.** The structure 1hcr is of the domain involved in ДНК sequence recognition (Figure 1f). It is a three-helix bundle flanked by short peptide chains at either end (about 50 residues). The second two  $\alpha$  helices form the HTH motif .

**Binding.** The full protein cooperatively binds as a homodimer at a 26 bp site. The recognition helix in the HTH motif is inserted in the major groove and surrounding helices make contacts with the phosphate backbone. The amino- and carboxy-terminal tails bind in adjacent minor grooves although their importance in sequence recognition is unknown.

#### 7. RAP1 family

**Function.** The RAP1 protein performs two functions. The first is the periodic binding of ДНК to regulate telomere length. Telomeres are nucleoprotein complexes found at the ends of eukaryotic chromosomes where the ДНК consists of a repeated array of short, species-specific sequence motifs. The second function is that of transcription regulation; RAP1 functions as an activator or repressor for a large number of genes.

**Structure.** RAP1 is a monomeric protein with two homologous domains and a carboxy-terminal tail (Figure 1g). Domain 1 (about 80 residues) contains a three-helix bundle and an amino-terminal tail, whereas domain 2 (about 80 residues) contains an additional fourth  $\alpha$  helix. In each, the second and third  $\alpha$  helices form the HTH motif. The two domains are connected by a 30 residue linker region and are positioned 8 bp apart. The carboxy-terminal tail is a 20 residue segment which emerges from domain 2 and folds back towards domain 1 .

**Binding.** The binding site is 16 bp long and shows a tandem repeat at an 8 base interval. The two domains bind in a similar fashion at opposite ends of the binding site; the recognition helices of the HTH motif are inserted in the major groove and the remaining  $\alpha$  helices contact the neighboring ДНК backbone. The amino-terminal and the linker regions interact with the minor groove and the carboxy-terminal tail interacts with the major groove as it folds back. The flexibility of the linker allows for slight variations in spacing between tandem repeats.

#### 8. Prd paired domain family

**Function.** The Prd paired domain is a functional domain found in a set of transcription regulatory proteins which are important in cell development.

**Structure.** The protein acts as a monomer with two structural domains (Figure 1h). The amino-terminal domain (about 70 residues) contains a short antiparallel  $\beta$  sheet and a  $\beta$  turn followed by a three-helix bundle and extended carboxy-terminal tail. The second and third  $\alpha$  helices in the bundle form an HTH motif. The carboxy-terminal domain (approximately 50 residues) also contains a three-helix bundle which has an HTH motif.

**Binding.** Prd proteins bind to 13-20 bp sites which share a common core sequence. The recognition helix in the HTH and the  $\beta$  turn of the amino-terminal domain make base contacts in the major and minor grooves respectively. The rest of the domain interacts with the ДНК backbone. The carboxy-terminal domain does not contact the ДНК, but domain structure and biochemical evidence suggest it does bind ДНК in certain family members (for example Pax proteins).

#### 9. Tc3 transposase family

**Function.** The structure contained in 1tc3 (Figure 1i) is of the ДНК-recognition domain found in the amino terminus of Tc3 transposase. The function of the enzymes is to move specific segments of ДНК from one position of the genome to another.

**Structure.** The domain (about 50 residues) contains a three-helix bundle and an amino-terminal tail (Figure 1i). The last two  $\alpha$  helices form the HTH motif .

**Binding.** Binding is to a 20 bp site. The recognition helix of the HTH motif is bound in the major groove and other  $\alpha$  helices make ДНК backbone contacts. The amino-terminal tail binds in an adjacent minor groove although the interactions are not thought to be specific.

#### 10. Trp repressor family

**Function.** The Trp repressor is involved in the regulation of tryptophan synthesis by binding three different operator sites. L-tryptophan acts as co-repressor.

**Structure.** Each subunit (about 100 residues) forms a six-helix bundle (Figure 1j). Helices 4 and 5 correspond to the HTH motif whereas the remaining four  $\alpha$  helices provide the dimerization interface. Tryptophan also binds the helical bundle .

**Binding.** Binding is to three related 16 bp operator sites which the protein binds in the presence of tryptophan. The HTH motifs are reoriented on binding of the co-repressor to enable ДНК-binding. The recognition helix is positioned in the major groove and most base contacts are made through a network of intermediate water molecules. Operator sites are symmetrical and also show approximate symmetry within the half-site, which leads to two alternative modes of binding. In the first, the dimer subunits bind each half-site symmetrically about the central base-pairs. This is similar to what is observed for the other prokaryotic HTH proteins. In the second, two dimers co-operatively bind to a single operator site in tandem. Dimers are staggered by 8 bp and rotated through 270° about the ДНК axis and the crystal structure 1trr (Figure 1j) displays a superhelix of dimers binding successive binding sites.

#### 11. Diphtheria tox repressor family

**Function.** The virulent phenotype of the pathogenic bacterium *Corynebacterium diphtheriae* is conferred by diphtheria toxin, whose expression is an adaptive response to low concentrations of iron. The expression of the toxin gene (*tox*) is regulated by the repressor diphtheria Tox, which is activated by transition metal ions.

**Structure.** Diphtheria tox is a 225-residue protein that binds as a dimer to ДНК. Each monomer consists of six helices and a short two-stranded  $\beta$  sheet, with helices 2 and 3 constituting the HTH motif (Figure 1k).

**Binding.** The ДНК interacts with two dimers bound to opposite sides of the *tox* operator, with each dimer interacting with two major groove regions. Together, the two HTH motifs (one in each dimer) bind a 24 bp sequence.

#### 12. Transcription factor TFIIB

**Function.** The transcription factor TFIIB is an essential part of the multiprotein transcription initiator complex that assembles on RNA polymerase II promoters. TFIIB binds a 7 bp region upstream of the TATA box called the B recognition box.

**Structure.** TFIIB is composed almost entirely of  $\alpha$  helices and is approximately 200 residues long (Figure 1l).

**Binding.** TFIIB binds ДНК in two places as a result of the nucleic acid distortion caused by the interaction of the TATA box-binding protein. The main interactions are due to a carboxy-terminal HTH motif binding ДНК in the major groove at the upstream site. The protein also binds ДНК in the minor groove at a downstream site using the amino terminus of a helix to contact the ДНК backbone.

#### 'Winged' HTH proteins

#### 13. Interferon regulatory factor family

**Function.** The family of interferon regulatory factor (IRF) transcription factors is important in the regulation of inter-ferons in response to infection by virus and in the regulation of interferon-inducible genes.

**Structure.** The IRF family is characterized by a unique 'tryptophan cluster' ДНК-binding region of five tryptophan residues. The protein binds as a monomer with a HTH motif binding

ДНК through three of the five conserved tryptophans. The IRF ДНК-binding region has an  $\alpha/\beta$  architecture consisting of a cluster of three  $\alpha$  helices flanked on one side by a mixed four-stranded  $\beta$  sheet (Figure 1m).

**Binding.** Helices 2 and 3 comprise the HTH motif, with helix 3 lying in the ДНК major groove. Contacts to bases within the major groove are localized to a GAAA core sequence within a 13 bp ДНК element in the interferon promoter.

#### 14. Catabolite gene activator (CAP) family

**Function.** CAP is a cAMP-dependent transcription regulator. A rise in cAMP concentration leads to increased affinity of CAP for catabolite-sensitive operons.

**Structure.** The protein functions as a homodimer, and each subunit comprises a two-domain structure (Figure 1n). The carboxy-terminal domain (about 60 residues) mainly consists of a three-helix bundle with the second two  $\alpha$  helices forming the HTH motif. The domain contains a small  $\beta$  sheet that also contributes to ДНК binding. The larger amino-terminal domain (approximately 130 residues) has an extensive  $\beta$  sheet that mediates cAMP binding, and a long  $\alpha$  helix that forms the dimer interface.

**Binding.** The consensus binding sequence is a symmetric 22 bp site. Binding by the recognition helix of the HTH motif in the major groove induces a sharp, highly localized bend in the ДНК and additional contacts with the phosphate backbone are made by the  $\beta$  strands from the same domain.

#### 15. Transcription factor family

##### Heat-shock and E2F/DP transcription factors

**Function.** The protein 3hts (Figure 1o) recognizes the promoters of the heat-shock protein genes through upstream ДНК sequences (heat-shock elements, HSEs). An HSE consists of alternating, inverted repeats of the sequence nGAAn, where n can be any nucleotide. The E2F and DP protein families form heterodimeric transcription factors that have a central role in the expression of cell-cycle-regulated genes and recognize a c/gGCGCg/c sequence.

**Structure.** The ДНК-binding domains of these proteins have a 'winged' HTH fold - that is, a three-helix bundle capped by an antiparallel  $\beta$  sheet. Helices 2 and 3 constitute the HTH motif.

**Binding.** The third helix of the HTH is docked into the major groove. The ДНК-binding domain makes additional contacts to the ДНК through the amino terminus of the first helix and the turn of the HTH motif. The only other HTH fold that contacts the ДНК with the residues of the turn is the Ets family.

#### 16. Ets domain family

**Function.** The Ets family of transcription factors, of which there are now about 35 members, regulate gene expression during growth and development. They share a conserved domain of around 85 amino acids which binds as a monomer to the ДНК sequence 5'-C/AGGAA/T-3'.

**Structure.** The 'winged' HTH motif interacts with a 10 bp region of duplex ДНК that takes up a uniform curve of  $8^\circ$  (Figure 1p).

**Binding.** The domain contacts the ДНК by a loop-helix-loop architecture, the turn of the HTH motif and the loop at the end of helix 1 before the  $\beta$  sheet contacting the ДНК backbone.

##### Group II: zinc-coordinating proteins

#### 17. $\beta\beta\alpha$ zinc-finger family

**Function.** The  $\beta\beta\alpha$  zinc-finger proteins constitute the largest individual family in this group. The ДНК-binding motif is found in many transcription regulators and more than a thousand distinct motifs have been identified through sequence analysis.

**Structure.** The structure of the finger is characterized by a short two-stranded antiparallel

$\beta$  sheet followed by an  $\alpha$  helix (Figure 2a) and a single zinc ion bound by two pairs of conserved histidine and cysteine residues situated in the  $\alpha$  helix and second  $\beta$  strand. Proteins generally contain multiple copies of fingers in a single peptide chain which wrap round the ДНК along the major groove in a spiral manner.

**Binding.** The recognition pattern of the probe  $\alpha$  helix has been well characterized; each finger binds adjacent 3 bp sub-sites on the ДНК using amino acids at positions -1, 2, 3 and 6 relative to the start of the  $\alpha$  helix, -1 being the residue position preceding the helix. Although exceptions to this rule have been observed in specific examples, experiments have shown that by altering the amino-acid types at the key positions, different subsite sequences are recognized, suggesting that these residue positions are usually sufficient for specific binding. By varying the number of fingers used in a protein chain, this relatively simple motif allows recognition of a wide range of binding sites with different degrees of specificity. For example, a protein with five fingers is expected to bind a site very selectively, whereas a protein with only a single finger would bind a wide range of sites containing the required 3 bp sequence. However, the structure of the human glioblastoma protein suggests that binding is not always straightforward; of the five fingers in the structure, one does not contact the ДНК at all and only two appear to make specific contacts with bases. As described earlier, the protein subunits in this study have been split into distinct domains, each containing a single zinc-finger motif. The pairwise sequence identities of the aligned domains are all high, ranging from 73% (for example, human zinc-finger protein, 1udbA1, and *Drosophila* tramtrack protein, 2drpA1) to 100% (for example, mouse Zif268 protein, 1aayA1, and artificial protein, 1mey). All domains are structurally very similar, returning SSAP scores of over 90.

#### 18. Hormone receptor family

**Function.** Members of the hormone receptor family translocate from the cytoplasm to the nucleus and regulate transcription at ДНК sequences called hormone response elements on binding of steroid and other hormones.

**Structure.** Hormone receptors function as homo- or hetero-dimers and each monomer typically consists of a ligand-binding, a ДНК-binding and a transcription regulatory domain (Figure 2b). The zinc-coordinating motif is found in the ДНК-binding domain and is characterized by two antiparallel  $\alpha$  helices capped by loops at their amino-terminal ends. Each helix-loop pair coordinates a single zinc ion using four conserved cysteines. The two  $\alpha$  helices lie approximately at right angles to each other; the first is inserted in the ДНК major groove to provide interactions with bases whereas the loops and the second  $\alpha$  helix contact the ДНК backbone. The ДНК-binding domain alone is sufficient for dimerization, the interface being formed by the loops leading into the second  $\alpha$  helix.

**Binding.** All receptor subunits bind to one of two half-site sequences, 5'-AGAACA-3' or 5'-AGGTCA-3'. A hormone-response element contains two half-sites and the identity of the response element is determined by the sequences that are present, the relative orientation between them (either symmetric or palindromic) and the spacing between them (between 3 and 6 bp). Thus recognition of the target sequence by the whole hormone receptor depends on read-out of half-site sequences by each subunit and the structure of the homo-or heterodimeric protein. The sequences of all subunits in the current dataset are very similar (sequence identities > 90%) except for the ДНК-binding domain of the thyroid hormone receptor (for example, 1bsx), which has two extra helices in the carboxy-terminal tail. The structures are all very similar with pairwise SSAP scores of over 90.

#### 19. Loop-sheet-helix family

**Function.** The loop-sheet-helix zinc-binding motif is represented solely by the ДНК-binding region of p53, a transcriptional activator implicated in tumor suppression.

**Structure.** As the name indicates, the ДНК-binding domain consists of a loop leading out of the main body of the protein, followed by a small  $\beta$  sheet, an  $\alpha$  helix and then another loop that leads back into the protein (Figure 2c). The zinc ion is coordinated by three cysteines and a histidine in the two loop regions.

**Binding.** Base contacts are supplied by the  $\alpha$  helix in the ДНК major groove and by the loops in the minor groove, although the latter are not thought to confer much specificity. The protein functions as a tetramer, with each subunit contacting a separate 5 bp recognition sequence positioned one after another. All intersubunit interactions are made by regions outside the ДНК-binding motif.

#### 20. Gal4-type family

**Function.** The final zinc-coordinating family contains only the Gal4 protein. It is a transcriptional regulator of galactose-induced genes and its zinc-coordinating motif has so far only been identified in proteins from *Saccharomyces cerevisiae*.

**Structure.** The motif consists of a pair of  $\alpha$  helices that coordinate two zinc ions through six cysteine residues, where two of the cysteines are shared by both metal atoms (Figure 2d).

**Binding.** The first  $\alpha$  helix is presented in the ДHK major groove for binding with bases, and backbone interactions are made by the second  $\alpha$  helix. Gal4 functions as a homodimer and the dimerization interface is located outside the zinc-coordinating motif.

#### Group III: zipper-type proteins

##### 21. Leucine zipper family

**Function.** The leucine zipper family consists of the yeast GCN4 proteins that bind promoter regions of genes encoding enzymes involved in amino-acid biosynthesis, and the Fos-Jun heterodimer, which activates the expression of many immune-response genes.

**Structure.** The structure of the zipper-type proteins may be split into two parts: the dimerization and ДHK-binding regions. As shown in (Figure 3a), each subunit in the leucine zipper protein consists of a single  $\alpha$  helix about 60 amino acids long. Dimerization is mediated through the formation of a coiled coil by a section of 30 amino acids at the carboxy-terminal end of each helix. The segment, known as the zipper region, consists of leucine or a similar hydrophobic amino acid every eight residue positions, roughly every two turns of the  $\alpha$  helix. Corresponding side chains from each subunit mediate hydrophobic contacts at the interface through side-by-side packing. The ДHK-binding region, also known as the basic region, is found in the amino terminus, and for the leucine zipper proteins, the binding segment is a direct extension of the dimerization region.

**Binding.** The  $\alpha$  helices of the two subunits diverge from the coiled coil and enter the ДHK major groove in opposing directions, each binding to half of the target sequence.

##### 22. Helix-loop-helix family

**Function.** The helix-loop-helix proteins are transcription factors that control the expression of a wide range of genes involved in differentiation and development.

**Structure.** As the name suggests, helix-loop-helix proteins are a modification of the continuous  $\alpha$  helices of the leucine zipper proteins in which the ДHK-binding and dimerization regions are separated by a loop, resulting in a four-helix bundle (Figure 3b).

**Binding.** Like the leucine zippers, the dimerization helices interact with each other in a coiled-coil arrangement and the ДHK-binding helices are inserted into the ДHK major groove. By separating the two segments, more flexibility is allowed in positioning the probe helices on binding nucleic.

The helix-loop-helix family is represented by the mouse and human forms of Max, Srebp-1, mouse MyoD and human USF proteins. Sequence identities range from 66% (Max protein, 1an2A, and USF protein, 1an4A) to 97% (mouse Max protein, 1an2A, and human Max protein, 1hloA) and with the exception of the MyoD (1mdyA) and USF (1an4A) protein pair (pairwise SSAP score 70), SSAP scores are above 80. Structural differences between proteins mainly arise from the variation in lengths and positioning of the loops.

#### Group IV: Other $\alpha$ -helix proteins

##### 23. Papillomavirus-1 E2 family

**Function.** This family has a single member, the papillo-mavirus-1 E2 protein, which uses a probe helix as part of the ДHK-recognition domain. The protein is a viral transcription regulator that acts at all viral promoters and also functions as a viral replication initiator.

**Structure.** The ДHK-binding region of the E2 protein (Figure 4a) is about 85 residues long and consists of four  $\beta$  strands and two interstrand  $\alpha$  helices. Two subunits combine to form an eight-strand  $\beta$ -barrel, which provides the interface for the resulting homodimer.

**Binding.** The larger  $\alpha$  helix from each subunit is symmetrically inserted in the ДHK major

groove making base and backbone contacts. Additional interactions to the backbone are provided by interstrand loops.

#### 24. Histone family

**Function.** ДНК in chromatin is organized in arrays of nucleosomes. The nucleosome, in its role as the principal packaging element of ДНК within the nucleus, is the primary determinant of ДНК accessibility.

**Structure.** Two copies of each of four histone proteins are assembled into an octamer that has 145-147 bp of ДНК wrapped in a superhelix around it to form a nucleosome core.

**Binding.** The protein octamer is divided into four 'histone-fold' dimers, each dimer being defined by H3-H4 and H2A-H2B histone pairs. The central histone-fold domains of all four core histone proteins share a highly similar structural motif constructed from three  $\alpha$  helices connected by two loops. The two H3-H4 pairs interact through a four-helix bundle formed only from the two H3 histone folds to define the H3-H4 tetramer. Each H2A-H2B pair interacts with this tetramer through a second, homologous four-helix bundle between H2B and H4 histone folds. The histone-fold regions of each tetramer bind to the center of the ДНК, which is wrapped into a superhelix. Further  $\alpha$  helices and coil elements extend from the histone-fold regions and are also an integral part of the core protein within the confines of the ДНК superhelix.

#### 25. EBNA1 protein (Epstein-Barr nuclear antigen 1)

**Function.** EBNA1 binds to four recognition sites in the origin of latent ДНК replication of Epstein-Barr virus and activates latent-phase replication of the viral genomes.

**Structure.** EBNA1 comprises two domains (Figure 4c, a flanking and a core domain (which is structurally homologous to the complete ДНК-binding domain of the bovine papilloma virus E2 protein) and binds ДНК as a dimer.

**Binding.** The flanking domain, which includes a helix that projects into the major groove and an extended chain that travels along the minor groove, makes all of the sequence-determining contacts with the ДНК. The core domain makes no direct contacts with the ДНК bases.

#### 26. Skn-1

**Function.** Skn-1 is a developmental transcription factor that specifies mesoderm in *Caenorhabditis elegans*.

**Structure.** Skn-1 consists of a compact four-helix unit with one helix more than twice as long as any of the others (Figure 4d).

**Binding.** It binds as a monomer and binds ДНК at two contact points. At the carboxy terminus, the longest helix extends from the domain to occupy the major groove of ДНК in a manner similar to zipper proteins. Skn-1, however, lacks the leucine zipper found in all zipper. Additional contacts with the ДНК are made by a short basic segment at the amino terminus of the domain, reminiscent of the 'homeodomain arm'.

#### 27. Cre recombinase family

**Function.:** Cre recombinase catalyzes a site-specific recombination reaction between two 34-bp *loxA* and *loxP* sites in bacteriophage  $\lambda$ .

**Structure.** Cre is a 320-residue protein and folds into two distinct domains that are separated by a short linker. The amino-terminal domain contains five helices and the large carboxy-terminal domain is primarily  $\alpha$  helical with a small  $\beta$  sheet packing against a nine-helix domain (Figure 4e).

**Binding.** The protein binds ДНК as a dimer, each monomer binding the outermost 15 bp of one *lox* half-site. The amino- and carboxy-terminal domains form a clamp around the half-sites making extensive contacts with both major and minor grooves. Helices 2 and 4 of the amino-terminal domain cross each other, both contacting the major groove of the *lox* half-site. The interface of ДНК with the carboxy-terminal domain is complex, involving the entire face of the domain, with both helices and connecting loops interacting with the major and minor

grooves and the ДНК backbone.

### 28. High-mobility group family

**Function.** The high-mobility group (HMG) chromosomal proteins, which are common to all eukaryotes, bind ДНК in a non-sequence-specific fashion to promote chromatin function and gene regulation. They interact directly with nucleosomes and are believed to be modulators of chromatin structure. They are also important in activating a number of regulators of gene expression, including p53, Hox transcription factors and steroid hormone receptors, by increasing their affinity for ДНК.

**Structure.** Chromosomal HMG proteins have a global fold of three helices stabilized in an 'L-shaped' configuration by two hydrophobic cores (Figure 4f).

**Binding.** The HMG domain binds to an AT-rich ДНК sequence using a large surface on the concave face of the protein, to bind the minor groove of the ДНК. This bends the ДНК helix axis away from the site of contact. The first and second helices contact the ДНК, their amino termini fitting into the minor groove, whereas helix 3 is primarily exposed to solvent. Partial intercalation of aliphatic and aromatic residues in helix 2 occurs in the minor groove.

### 29. MADS-box family

**Function.** The MADS-box motif is found in various ДНК-binding proteins, commonly transcription factors, and specifies ДНК binding, dimerization and interaction with accessory factors.

**Structure.** MADS proteins bind ДНК as dimers as part of a larger cooperative ДНК-binding complex containing other ДНК-binding proteins. The MADS domain is a 56-residue motif consisting of a pair of antiparallel coiled-coil  $\alpha$  helices packed against an antiparallel two-stranded  $\beta$  sheet. This  $\beta$  sheet of the motif is also involved in interprotein interactions with other accessory proteins.

**Binding.** MADS dimerization occurs along the extensive flat side of the monomer involving the helices and  $\beta$  sheet. The MADS protein shown here, MCM-1 (Figure 4g), interacts with ДНК predominantly with its long  $\alpha$  helices located nearly parallel to the minor groove at the center of the binding site. These  $\alpha$  helices extend into the major groove on either side of the dyad; direct contacts made within the major groove and along the phosphate backbone cause the ДНК to bend around the MADS box. The amino-terminal strand of the MADS region (before the first helix of the MADS motif) often passes over and interacts with the ДНК backbone.

### Group V: $\beta$ -sheet proteins

#### 30. TATA box-binding family

This group, which only contains the TATA box-binding protein family, is characterized by the use a large  $\beta$ -sheet structures to bind the ДНК (Figure 5).

**Function.** TATA box-binding proteins are an essential component of the multiprotein transcription initiator complex that assembles on promoters bound by RNA polymerase II.

**Structure.** Although they are single-chain molecules, their structures are generally considered to consist of two pseudoidentical domains. A ten-stranded antiparallel  $\beta$  sheet joins the domains.

**Binding.** The  $\beta$  sheet covers the ДНК minor groove and creates two substantial kinks away from the main body of the protein, by intercalating phenylalanine side chains from either end of the sheet.

The family is represented by *Pyrococcus woesei*, *Saccharomyces cerevisiae* and human forms of the protein. Unsurprisingly, both sequence and structural alignments of the various subunits yield very high scores (> 90% and 90 or more respectively).

### Group VI: $\beta$ -hairpin/ribbon proteins

#### 31. MetJ repressor family

**Function.** Transcriptional regulator of the expression of methionine biosynthetic enzymes

in *E. coli*.

**Structure.** The MetJ repressor binds ДНК as a dimer (Figure 6a), each subunit comprising a helical bundle and a single  $\beta$  strand; the strands from each subunit form the antiparallel sheet for ДНК-binding (colored red).

**Binding.:** The two  $\beta$  strands fit into the major groove and do not alter the ДНК structure significantly on binding. They lie flat against the base of the groove and interactions are only made from one face of the sheet. Supporting backbone contacts are made by the surrounding helices and the amino-terminal loop regions.

### 32. Tus replication terminator family

**Function.** Tus protein terminates replication of ДНК in *E. coli*.

**Structure.** The protein consists of two  $\alpha$ -helical bundles at the amino and carboxy termini, connected by a large  $\beta$ -sheet region and binds ДНК as a monomer.

**Binding.** The ДНК-binding region of the Tus family is made of four antiparallel  $\beta$  strands (colored red in Figure 6b) which links the amino- and carboxy-terminal domains and produces a large central cleft in the protein. The ДНК is bound in this cleft, with the interdomain  $\beta$  strands contacting bases in the major groove. ДНК backbone contacts are provided by the whole protein. The  $\beta$  strands are positioned almost perpendicular to the base edges in the groove, enabling contacts from amino acids that expose their side chains on either face of the sheet.

### 33. Integration host factor family

**Function.** Integration host factor (IHF) is a small heterodimeric protein that specifically binds to ДНК and functions as an architectural factor in many cellular processes in prokaryotes.

**Structure.** The protein is a heterodimer of two related subunits each made of three helices and a two-stranded  $\beta$  sheet.

**Binding.** In contrast to the two families above, the integration host factor forces an enormous distortion in the ДНК by inserting a  $\beta$  hairpin from each subunit in the minor groove (red in Figure 6c). As seen in the TATA box-binding family, the protein produces kinks by intercalating side chains between base steps at the edges of the binding sites. The intercalating prolines are found at the tips of the  $\beta$  hairpins that extend from the protein towards the other side of the ДНК. The nucleic acid is bent towards the main body of the protein and the deformation is stabilized by contacts with the phosphate groups .

### 34. T-domain family

**Function.** The T domain (Figure 6d) is an approximately 180-residue homodimeric domain found in transcriptional regulators for genes essential in tissue specification, morphogenesis and organogenesis.

**Structure.** Each subunit consists of a seven-strand antiparallel  $\beta$  barrel; one opening of this barrel forms a dimer interface with the equivalent segment of the other subunit while the other end points towards the ДНК.

**Binding.** Two  $\beta$  strands protrude from the barrel, one of which extends into the ДНК major groove. The probe helix is situated in a three-helix bundle in the carboxy-terminal tail. In contrast to many protein families, the  $\alpha$  helix binds base and backbone groups from the ДНК minor groove .

### 35. Hyperthermophile chromosomal proteins

**Function.** These proteins are found in hyperthermophilic archaeobacteria and have high thermal, acid and chemical stability. They bind ДНК without marked sequence preference and increase the  $T_m$  of ДНК by about 40°C.

**Structure.** The proteins consist of an incomplete five-stranded  $\beta$ -barrel capped by an  $\alpha$  helix abutting three  $\beta$  strands (Figure 6e).

**Binding.** The proteins bind the minor groove with the three-stranded  $\beta$  sheet causing the ДНК to kink severely. The kink results from the intercalation of specific hydrophobic side chains into the ДНК structure, but without causing any significant distortion of the protein structure relative to the uncomplexed protein in solution.

### 36. Arc repressor

**Function.** Transcription of the *ant* gene during lytic growth of bacteriophage P22 is regulated by the cooperative binding of two Arc repressor dimers to a 21-bp operator site.

**Structure.** Arc is a small (about 100 residues), homodimeric repressor of the ribbon-helix-helix family of transcription factors. Each monomer consists of a pair of helices connected by an antiparallel  $\beta$  sheet (Figure 6f).

**Binding.** Each Arc dimer uses the  $\beta$  sheet to recognize bases in the major groove and the amino termini of the second helix in each pair contact the ДНК backbone.

### Group VII: other

### 37. Rel homology region family

**Function.** The Rel homology region is found in the amino terminus of proteins that act at the  $\kappa$  B ДНК recognition site, and mediates ДНК binding, dimerization and nuclear localization (Figure 7a). Proteins that contain the region act as transcription regulators for genes commonly involved in cellular defense and differentiation. The carboxy-terminal domains located outside the region are variable between proteins.

**Structure.** The Rel homology region binds symmetrically as a homo- or hetero-dimer. Each subunit (of about 300 residues) has two distinct domains, both consisting of a  $\beta$  sandwich.

**Binding.** Interactions in the ДНК major groove are made along the whole length of the 10 bp site using a total of ten interstrand loops .

### 38. STAT protein family

**Function.** STATs are a family of eukaryotic transcription factors that mediate the response to a large number of cytokines and growth factors. Upon activation by cell-surface receptors or their associated kinases, Stat proteins dimerize, translocate to the nucleus and bind to specific promoter sequences.

**Structure.** STAT proteins are between 750 and 850 residues long and bind as dimers to ДНК target sites with a 9 bp consensus sequence, TTCCGGGAA. Each monomer is composed of four domains: an amino-terminal four-helix bundle, an eight-stranded  $\beta$  barrel (residues 321-465), a helix-loop-helix 'connector' domain (residues 466-585) and an SH2 domain.

**Binding.** The STAT homodimer grips the ДНК like a pair of pliers (Figure 7b). The monomers are held together by the carboxy-terminal SH2 domains, and the large four-helix bundle domains form the 'handles' of the pliers. The ДНК is almost entirely enclosed by the protein dimer, and contacts the loops from the  $\beta$  barrel and the connector domains.

### Group VIII: enzymes

### 39. Methyltransferase family

**Function.** The methyltransferase enzyme is represented by a single homologous family. The protein catalyzes the transfer of a methyl group from S-adenosyl-L-methionine to the C5 position of cytosine. In prokaryotes the reaction is most commonly found in the protection of the ДНК from restriction enzymes. In eukaryotes, however, ДНК methylation is implicated in a wider range of cellular processes including transcriptional regulation, ДНК repair, developmental regulation and chromatin organization. The current dataset only includes the prokaryotic *HhaI* methyltransferase (for example, 4mht).

**Structure.:** The protein functions as a monomer (about 320 residues) containing two domains that are separated by a large ДНК-binding cleft (Figure 8a). The catalytic domain (about 220 residues) consists of a seven-stranded  $\beta$  sheet flanked by a total of five  $\alpha$  helices on either side. This domain contains the cofactor-binding site and the active sites. The ДНК-

recognition domain (about 100 residues) comprises five antiparallel strands that form a twisted  $\beta$  sheet.

**Binding.** The protein preferentially binds the sequence 5'-GCGC-3' with the first cytosine base methylated in the enzyme reaction. The ДНК is bound in the protein cleft so that the major groove faces the recognition domain and the minor groove faces the catalytic domain. The 4 bp in the target sequence are contacted from the major groove using two glycine-rich interstrand loops, and the substrate cytosine is flipped out of the ДНК helix into the catalytic domain. The ДНК structure is underwound and the base-pairing is rearranged over 3 bp either side of the substrate base. The three structures in the family all have identical sequences and return high pairwise SSAP scores ( $> 90$ ).

#### 40-44. Endonucleases

Seven endonuclease families are represented in the current dataset. The *FokI* family also belongs to the HTH group and has already been described. Figure 8b-8f display MolScript diagrams for representative structures of all the families, viewed parallel and perpendicular to the ДНК axis. *EcoRV*, *PvuII*, *EcoRI* and *BamHI* (1rva, 1piv, 1eri and 1bhm, respectively) are type II restriction endonucleases that recognize ДНК sites of 6 bp in length and cleave the phosphate backbone at precise positions within the target sequence. Although there is little sequence similarity between the four protein types, their U-shaped homodimeric structures display some very common features.

The subunits of *PvuII* (about 140 residues per subunit) and *EcoRV* (approximately 240 residues per subunit) may be divided into three segments: the amino-terminal dimerization region, the core catalytic region and the carboxy-terminal ДНК-recognition region (Figure 8b, 8c). The catalytic regions of both comprise a five- or six-stranded mixed parallel/antiparallel  $\beta$  sheet (colored blue), which forms part of the cavity base. Most of the ДНК-recognition segments extend from the carboxy-terminal end of the catalytic region (red). In *PvuII*, the region comprises two parallel  $\alpha$  helices and in *EcoRV*, a mixture of  $\alpha$  helices and  $\beta$  strands. Both proteins approach the minor groove, and the ДНК-recognition regions reach around the side of the ДНК to contact bases in the major groove using a pair of loops. The dimerization regions of the two proteins are very different (colored green) and complete the base of the cavity.

The catalytic (or core) regions of endonucleases *EcoRI* (about 250 residues per monomer) and *BamHI* (about 200 residues per monomer) also consist of five-stranded parallel/antiparallel  $\beta$  sheets (Figure 8d, 8e). The positioning of the sheets is different from *EcoRV* and *PvuII*, and they form the sides of the cavities. Included in the core region of both proteins are two  $\alpha$  helices that pack against their counterparts in the other subunit to form a four-helix bundle at the base of the cavity. *EcoRI* and *BamHI* both approach the ДНК and make most of the base contacts from the major groove, although the method of sequence recognition greatly differ. *EcoRI* uses an extra set of interstrand loops and strands that follow the major groove towards the outer edges of the target sequence from the center (green in Figure 8d). *BamHI* lacks these extra regions and uses the amino-terminal end of the helical bundle for binding.

#### 44. Endonuclease V

This protein (for example 1vas) catalyzes the first step in the pyrimidine-specific base-excision repair pathway. In contrast to the type II enzymes described above, endonuclease V functions as a monomer (about 130 residues) whose structure comprises a four-helix bundle arranged to form a concave surface in which the ДНК is bound (Figure 8f). Binding is centered on a damaged pyrimidine dimer; most of the interactions are to the ДНК backbone, and the only base contacts are made to the central adenine which is flipped out of the ДНК helix into a cavity on the protein surface.

#### 45. DNase I

**Function.** DNase I is an endonuclease that degrades double-stranded ДНК in a non-specific but sequence-dependent manner. Its function is dependent on the presence of divalent cations such as  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$  and  $\text{Mn}^{2+}$ .

**Structure.** DNase I is an  $\alpha,\beta$  protein with two six-stranded  $\beta$ -pleated sheets packed against each other forming the core of a 'sandwich'-type structure. The two predominantly antiparallel  $\beta$  sheets are flanked by three longer  $\alpha$  helices and extensive loop regions.

**Binding.** DNase I binds in the minor groove of the ДНК duplex with an exposed loop region forming contacts in and along both sides of the minor groove and extending over a total

of 6 bp (Figure 8g). As a consequence of DNase I binding, the minor groove opens by about 3 Å and the duplex bends towards the major groove by about 20°.

#### 46. ДНК mismatch endonuclease

**Function.** In *E. coli*, the enzyme recognizes a TG mismatched base pair, generated after spontaneous deamination of methylated cytosines, and cleaves the phosphate backbone on the 5' side of the thymine.

**Structure.** The protein contains three helices surrounding a β sheet, with one other helix used to intercalate the ДНК.

**Binding.** Three aromatic residues from one helix intercalate into the major groove of the ДНК to strikingly deform the base pair stacking (Figure 8h).

#### 47-50. Polymerase group

Polymerases must provide sequence-independent interactions with their ДНК substrate, yet retain the specificity to distinguish correctly paired bases from mismatches. ДНК polymerases synthesize ДНК strands by catalyzing the stepwise addition of a deoxyribonucleotide to the 3'-OH end of a polynucleotide chain that is paired to a second, template strand. Four polymerases have been classified: Pol β, Pol I, Pol T7 and Pol RT (reverse transcriptase).

47. ДНК polymerase β (pol β); 48. ДНК polymerase I (pol I); 49. ДНК polymerase T7 (pol T7)

Pol β (Figure 8i) and Pol I (Figure 8j) have three structural domains that perform three separate functions, not only polymerizing the ДНК but editing and repairing it by 3'-5'- and 5'-3'-exonuclease activity respectively. T7 ДНК polymerase (Figure 8k) possesses no 5'-3'-exonuclease activity. For Pol I and T7, the larger carboxy-terminal domain has both the polymerase and 3'-5'-exonuclease activity with an α+β structure that can be likened to that of a right hand. A large cleft formed from a six-stranded antiparallel β sheet surrounded by α helices forms the 'palm' and binds the ДНК minor groove along with the 'thumb' region (Figure 8j, 8k). Extensive sequence-independent interactions exist in the minor groove. The major groove, with its sequence-specific pattern of hydrogen-bond donors and acceptors, which form the primary means of recognition for many sequence-specific ДНК-binding proteins, does not contact the protein and is solvent-accessible.

The smaller amino-terminal of Pol I has 5'-3'-exonuclease activity. It is folded into an αβ structure with a mixed β sheet of five strands.

#### 50. HIV reverse transcriptase

**Function.** Reverse transcriptases have two enzymatic activities: a ДНК polymerase that can copy either ДНК or RNA templates and an RNase H. The two crystal structures of HIV reverse transcriptase which have been solved are only of the polymerase region.

**Structure.** HIV-1 reverse transcriptase (Figure 8l) is a heterodimer consisting of p66 (about 550 residues) and p51 (about 430 residues), two subunits of α helices and β strands which share a common amino terminus. The p51 subunit corresponds to the polymerase domain of the p66 subunit. The carboxy terminus of p66 forms the RNase H domain.

**Binding.** Loops and helices of p66 make extensive interactions with the ДНК. P51 also binds but its interactions are mainly at the protein dimer interface with p66.

#### 51. Uracil-ДНК glycosylase

**Function.** Any uracil bases in ДНК, a result of either misincorporation or deamination of cytosine, are removed by uracil-ДНК glycosylase (UDG).

**Structure.** UDG is 225 residues long and contains a central four-stranded β-sheet region partly surrounded by eight α helices (Figure 8m).

**Binding.** Damaged ДНК binds to UDG near the carboxy-terminal end of its central four-stranded β sheet. Conserved UDG residues in loop regions contact the ДНК, with the loop

between sheet 4 and helix 8 inserting into the ДНК minor groove. A few contacts with the ДНК backbone are made by two helices.

### 52. 3-Methyladenine ДНК glycosylase

**Function.** ДНК N-glycosylases are base excision-repair proteins that locate and cleave damaged bases from ДНК as the first step in restoring the sequence.

**Structure.** The protein is 216 residues in length and is composed mainly of  $\beta$  strands (Figure 8n).

**Binding.** The enzyme intercalates into the minor groove of ДНК using two  $\beta$  strands, causing the damaged base to flip into the enzyme active site for base excision.

### 53. Homing endonuclease family

**Function.** Homing endonucleases are a diverse collection of proteins that are encoded by genes with mobile, self-splicing introns. These enzymes promote the movement of the ДНК sequences that encode them from one chromosome location to another; they do this by making a site-specific double-strand break at a target site in an allele that lacks the corresponding mobile intron.

**Structure.** The protein binds ДНК as a dimer and displays mixed  $\alpha\beta$  topology (Figure 8o). Each monomer contains three antiparallel  $\beta$  sheets flanked by two long  $\alpha$  helices, and a long carboxy-terminal tail that extends around the surface of the second subunit in the dimer and is stabilized by two bound zinc ions 15 Å apart.

**Binding.** The zinc-binding motifs are critical primarily for structural stabilization of the protein core and are not involved in ДНК binding. The primary sequence-specific contacts made to homing-site ДНК are from residues in the second  $\beta$  sheet of each enzyme monomer which contact the major groove of each half-site. Additional contacts are made in the center of the complex within the minor groove and with several phosphate groups in the cleavage site.

### 54. Topoisomerase I

**Function.** Topoisomerases I promote the relaxation of ДНК superhelical tension by introducing a transient single-stranded break in duplex ДНК and are vital for the processes of ДНК replication, transcription and recombination.

**Structure.** No crystal structure has been solved for the whole protein - only for the central core and the carboxy-terminal domains (592 residues; see Figure 8p). The central core domain is connected to the carboxy-terminal domain by a linker. This linker assumes a coiled-coil configuration and protrudes away from the remainder of the enzyme.

**Binding.** The enzyme completely surrounds the ДНК, contacting the backbone with loops and a  $\beta$  sheet binds in the major groove.

## References

- Harrison SC: **A structural taxonomy of ДНК-binding domains.** *Nature* 1991 **353**: 715-719
- Luisi BF: **ДНК-protein interaction at high resolution.** *In ДНК-Protein Structural Interactions. Edited by Lilley DMJ. New York: Oxford University Press, 1995 : 1-48*
- Frishman D, Mewes H-W: **PEDANTic genome analysis.** *Trends Genet* 1997 **13**: 415-416
- Bernstein FC, Koetzler TF, Williams GJB, Meyer EF, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M: **The Protein Data Bank: a computer-based achival file for macromolecular structures.** *J Mol Biol* 1977 **112**: 535-542
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN,

Bourne PE: **The Protein Data Bank.**  
*Nucleic Acids Res* 2000 **28**: 235-242

• Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin , Demeny T, Hsieh S-H, Srinivasan AR, Schneider B: **The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids.**  
*Biophys J* 1992 **63**: 751-759

• Sayle RA, Milner-White EJ: **RasMol - Biomolecular graphics for all.**  
*Trends Biochem Sci* 1995 **20**: 374-376

• Orengo CA, Taylor WR: **SSAP: sequential structure alignment program for protein structure comparison.**  
*Methods Enzymol* 1996 **266**: 617-635

• Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequences of two proteins.**  
*J Mol Biol* 1970 **48**: 443-453

• Orengo CA, Flores TP, Taylor WR, Thornton JM: **Identification and classification of protein fold families.**  
*Protein Eng* 1993 **6**: 485-500

• Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.**  
*Nucleic Acids Res* 1997 **24**: 4876-4882

.....



→ [К оглавлению](#) | [К титульной странице](#) | [К нормальному развитию](#)