



# БАЗЫ ЗНАНИЙ ПО МОЛЕКУЛЯРНОЙ БИОЛОГИИ

*М.П.Пономаренко, Ю.В.Пономаренко, А.С.Фролов, Н.Л.Подколотный,  
О.А.Подколотная, Д.Г.Воробьев, В.Г.Левицкий, Н.А.Колчанов*

Лаборатория теоретической генетики ИЦиГ СО РАН

Институт цитологии и генетики СО РАН  
630090, Новосибирск, просп. ак. Лаврентьева, 10  
тел.: (3832) 33–31–19, факс: (3832) 33–12–78  
email: pon@bionet.nsc.ru

## **Введение**

«Геном человека» является одним из крупнейших всемирных научных проектов. В рамках этого проекта аннотация геномной ДНК включает предсказание неизвестных генов, регуляторных районов и функциональных сайтов с целью планирования прецизионных экспериментов по их точной идентификации. Это требует применения знаний о строении генов, регуляторных районов, функциональных сайтов; молекулярных механизмах их работы, генных сетях, координации работы генов в процессах жизнедеятельности организмов. Массив даже самых необходимых таких знаний настолько велик и разнороден, что его невозможно ни «держать в голове», ни издать в виде «Руководства по неизвестным генам». Современные информационно-поисковые и вычислительные технологии позволяют накапливать гигантские массивы знаний в специальных базах знаний, позволяющих быстро находить необходимые знания в процессе решения конкретных исследовательских задач.

В лаборатории теоретической генетики ИЦиГ СО РАН осуществляется создание баз знаний по молекулярной биологии для обеспечения полного цикла аннотации геномной ДНК, включая накопление экспериментальных данных; их анализ; документирование закономерностей, выявленных в результате этого анализа; генерацию компьютерных программ, использующих эти закономерности для аннотации геномной ДНК, а также (что является уникальным свойством баз знаний) объяснение результатов аннотации геномной ДНК вплоть до указания экспериментальных данных, на основании которых были созданы компьютерные программы для получения этих результатов.

## **Архитектура системы**

Место баз знаний по молекулярной биологии в компьютерной системе GeneExpress схематически показано на рисунке 1.

Можно видеть, что посредством баз знаний осуществляется взаимосвязь между известными экспериментальными данными и компьютерными программами, предназначенными для планирования новых экспериментов. Это является следствием того, что в базах знаний документируются закономерности, выявленные в результате анализа определенных экспериментальных данных, и эти же закономерности будут использоваться некоторой компьютерной программой для решения исследовательских задач. В свою очередь осуществление посредством баз знаний взаимосвязи между компьютерными программами и экспериментальными данными позволяет обеспечить три новые исследовательские возможности для решения молекулярно-генетических задач. Прежде всего, результаты каждой программы могут быть объяснены исследователю-генетику путем указания тех экспериментальных данных, закономерности которых были использованы при получении этих результа-

тов. Кроме того, из сотен документированных в базе знаний компьютерных программ пользователь-генетик с помощью стандартных информационно-поисковых средств может быстро найти именно те программы, которые необходимы ему на каждом шаге его исследования. Наконец, сопоставляя между собой результаты распознавания нескольких функциональных сайтов, координированно регулирующих экспрессию определенной группы генов, пользователь-генетик может осуществлять комплексный анализ таких генов, ограниченный текущим состоянием известных экспериментальных данных.

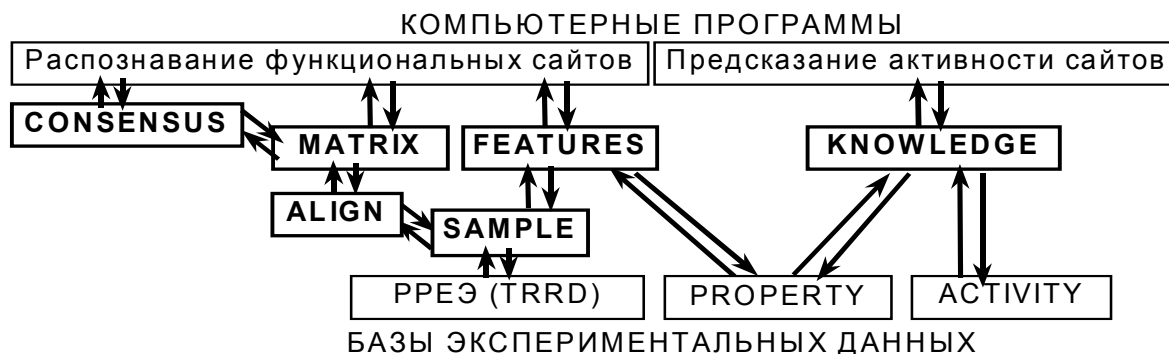


Рис. 1. Схема компьютерной системы GeneExpress [1]. Базы знаний (выделены жирным шрифтом) осуществляют взаимосвязь (стрелки) между экспериментальными данными и компьютерными программами для планирования новых экспериментов.

В данной работе создание и применение баз знаний по молекулярной биологии демонстрируются на примере решения двух конкретных биологических задач: распознавания функциональных сайтов и предсказания активности сайтов. Текущие версии представляемых баз знаний содержат: ACTIVITY – 49 знаний-программ для предсказания активности функциональных сигналов; ALIGN – 45 знаний-выборок выравненных последовательностей сайтов связывания транскрипционных факторов; FEATURES – 1402 знаний-программ распознавания сайтов по функционально важным конформационным и физико-химическим свойствам ДНК сайтов; MATRIX – 567 знаний-программ распознавания сайтов по частотам коротких «слов»-олигонуклеотидов; CONSENSUS – 66 знаний-программ распознавания сайтов по эволюционно-консервативным инвариантам этих сайтов. Эти базы знаний являются свободно доступными для молекулярных биологов-генетиков по сети ИНТЕРНЕТ, <http://wwwmgs.bionet.nsc.ru/systems/GeneExpress/>.

## Результаты и обсуждение

### 3.1. Базы знаний для распознавания функциональных сайтов ДНК

Элементарным шагом аннотации геномной ДНК является распознавание участков связывания ДНК (функциональных сайтов) с регуляторными белками. В качестве исходных экспериментальных данных для решения задачи распознавания сайтов используется база данных РРТЭ по регуляторным районам генов [2] (эта база данных известна также как «Transcription Regulatory Region Database, TRRD»). В ней описаны регуляторные районы более 650 генов, содержащие более 2500 сайтов связывания транскрипционных белковых факторов (рис. 1, внизу слева). В результате группировки этих сайтов по типам связывающих их белков создана база знаний SAMPLES, в текущей версии которой документировано более 50 выборок таких сайтов. В таблице 1 показана выборка сайтов связывания транскрипционного фактора YY-1 с регуляторными последовательностями генов эукариот. Можно видеть, что размеры экспериментально установленных участков связывания ДНК с белком значительно варьируют. Поэтому с помощью стандартного метода максимизации по-

тенциала Гиббса [3] для всех последовательностей ДНК сайта осуществляется множественное выравнивание и находится консенсус, которые документируются в базах знаний ALIGN и CONSENSUS соответственно.

Для выборки  $N$  выровненных последовательностей ДНК длины  $L$  исследуемого сайта  $\{S_n=(s_{n,i})\}_{i=1,L; n=1;N}$  в каждой его  $i$ -й позиции вычисляются частоты нуклеотидов и динуклеотидов в следующих 22-х алфавитах  $E_0=\{e_{01}=A, e_{02}=T, e_{03}=G, e_{04}=C\}$ ,  $E_1=\{e_{11}=W=A+T, e_{12}=S=G+C\}$ ,  $E_2=\{e_{21}=R=A+G, e_{22}=Y=T+C\}$ ,  $E_3=\{e_{31}=M=A+C, e_{32}=K=T+G\}$ ,  $E_4=\{e_{41}=WW, e_{42}=WS, e_{43}=SW, e_{44}=SS\}$ ,  $E_5=\{e_{51}=YY, e_{52}=YR, e_{53}=RY, e_{54}=RR\}$  и  $E_6=\{e_{61}=MM, e_{62}=MK, e_{63}=KM, e_{64}=KK\}$ :

$$F(e_{pj};i)=\{\sum_{n=1,N} \Pi_{q=1,Q} \delta(s_{n;i+q-1} \in e_{pj})\}/N; \quad (1)$$

здесь:  $p$  – номер алфавита,  $j$  – номер олигонуклеотида;  $Q$  – длина олигонуклеотида, равная «1» для  $E_0, E_1, E_2, E_3$  или «2» для  $E_4, E_5, E_6$ ;  $\delta(\text{ИСТИНА})=1$  и  $\delta(\text{ЛОЖЬ})=0$ .

Таблица 1

Сайты связывания ДНК с транскрипционным фактором YY-1

№	База данных: вход	Последовательность (множественное выравнивание)*
1	TRANSFAC [4]: R03174	cacatgggtgctgc AAAATGTC gcaaaacactcacg
2	TRANSFAC [4]: R03175	aacctcccgcttc AAAATGGA gacctgctgctc
3	TRANSFAC [4]: R00034	ggcccgacaccca AATATGGC gacggccggggccg
4	TRANSFAC [4]: R00466	acgcagatgtcct AATATGGA catcctgtgtaagg
5	TRANSFAC [4]: R03177	aaagtctccagaa AACCTAGA ggccacgggtcaag
6	TRANSFAC [4]: R00848	ttgattggagtc AAGATGGC cgatcagaaccaga
7	TRANSFAC [4]: R01149	ttcatgccttgc AAAATGGC gttacttaagctag
8	TRANSFAC [4]: R00284	caatggtcgaacc ATGATGGC agcggggataaaat
9	TRANSFAC [4]: R00688	aagagtatcggac CAGATTGA aaaccgaaagcggc
10	TRANSFAC [4]: R01225	ggccagctcgggt CAGTTAGT cacttctgcttaa
11	TRANSFAC [4]: R01219	tgatcagctcag AAGATGGC ggagggcctccaac
12	TRANSFAC [4]: R04077	tttccggctgga ACCATGGA ggctgttccgtaag
13	TRANSFAC [4]: R03002	ccggaagtgtca AAGATGGC tgctgtgatggctt
14	TRANSFAC [4]: R01329	actcaccgtaaaa CAGATGGC agccacctcgtagg
15	EMBL: K02061	tgatgccgtaaaa CAGATGGC agccacctcgtagg
16	TRANSFAC [4]: R01330	aagccgcgggcgg CGGATGGC cgccgatacagact
17	TRANSFAC [4]: R01833	aacaccagccgcc AAGATGGC cggggagcgagaaa
18	EMBL: K02929	aacaccagccgcc AAGATGGC cggggagtggagaaa
19	PPTЭ (SITES) [2]: 863	gggggctgcccc AAAATGGC cgggggctggggc
20	MEDLINE: 7969151	gtgatcctccgag CCAATGGC caccggtcgtcga
21	PPTЭ (SITES) [2]: 2010	tctgtccaccag CAAATGGC atttcagcattatt
22	PPTЭ (SITES) [2]: 1676	aaggagaatggga GAGATGGA tatcattttggaag
23	MEDLINE: 8821623	gaacaggctgagc AAGATGGG cgggacttccgttg
24	PPTЭ (GENES) [2]: Hs:BG	caaatgtaagcaa TAGATGGC tctgcctgacttt
25	MEDLINE: 8887668	gcccggagaatac AAAAAGGC acctgacggccgtc
26	PPTЭ (SITES) [2]: 1112	ttgcacaggaat GACATGGT gggactttcccag
27	MEDLINE: 2507312	agtaatatcaaac AAGATGGA ggtggggtatcatt

\* Экспериментально установленные участки связывания ДНК с белком подчеркнуты; консенсус сайтов AAGATGGC обозначен заглавными буквами.

Для каждого алфавита  $E_p$  по частотам  $F(e_{p;j})$  оценивается частное сходство  $D_p(S)$  произвольной последовательности  $S$  с рассматриваемым функциональным сайтом:

$$D_p(S) = \left[ \left\{ \sum_{i=1, L-Q+1} \sum_{j=1, J} F(e_{p;j}) \times \prod_{q=1, Q} \delta(s_{n;i+q-1} \in e_{p;j}) \right\} - \alpha_p \right] / \beta_p; \quad (2)$$

здесь:  $J$  – размер алфавита, равный «4» для  $E_0, E_4, E_5, E_6$ , или «2» для  $E_1, E_2, E_3$ ;  $\alpha_p$  и  $\beta_p$  – нормирующие коэффициенты, удовлетворяющие условию: «среднее значение  $D_p$  для всех сайтов  $\text{Mean}_{\text{SITE}}(D_p)=1$ , для случайных последовательностей ДНК  $\text{Mean}_{\text{RAND}}(D_p)=-1$ ». Интегральной оценкой сходства является среднее значение этих частных оценок:

$$D(S) = \sum_{p=0,6} D_p(S) / 7, \quad (3)$$

с распознающим правилом: « $D(S) > 0$ » означает « $S$  – рассматриваемый сайт».

Матрицы частот нуклеотидов и динуклеотидов документируются в базе знаний MATRIX (рис. 1). Для рассматриваемого сайта связывания транскрипционного фактора YY-1 на рисунке 2 показаны матрицы частот нуклеотидов и динуклеотидов. Из рисунка 2,а можно видеть, что достоверно частые нуклеотиды в позициях с 1 по 7 могут быть включены в консенсус сайта YY-1, тогда как в позициях 0, 8, 9 и 10 частоты нуклеотидов не превышают 95%-границу их достоверности. Рассмотрение частот динуклеотидов (рис. 2,б) позволяет уточнить консенсус сайта и включить позицию «0» в консенсус сайтов, так как в этой позиции имеется достоверно частый динуклеотид MM. Консенсус сайтов YY-1 «AAGATGGC» документирован в базе знаний CONSENSUS.

Рассмотренное построение консенсуса сайтов YY-1 свидетельствует о том, что учет частот динуклеотидов (рис. 2,б) уточняет результаты анализа нуклеотидных частот (рис. 2,а). Действительно, частоты динуклеотидов содержат информацию о предпочтительных «ближайших соседях», которой нет в частотах нуклеотидов.

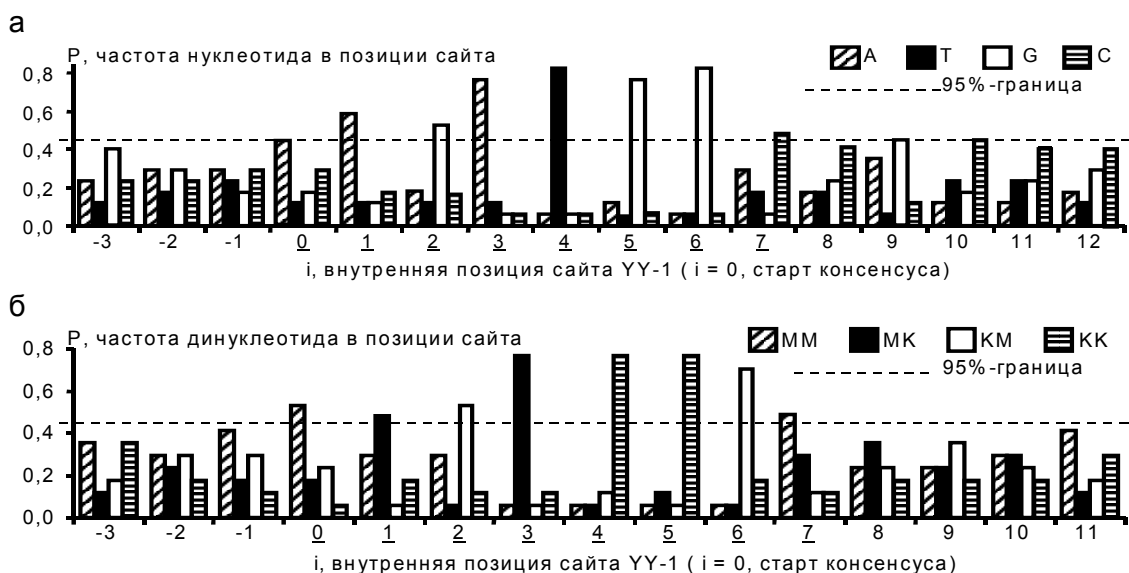


Рис. 2. Примеры информации о сайтах YY-1, документированной в базе знаний MATRIX: а – матрица частот нуклеотидов; б – матрица частот динуклеотидов (i=0 – старт консенсуса сайтов YY-1; W=A/T, S=G/C, R=A/G, Y=T/C, M=A/C и K=T/G). Разрывная линия – 95%-граница достоверно высокой частоты встречаемости.

На рисунке 3 показана оценка точности усредненного метода распознавания сайтов YY-1 по частотам нуклеотидов и динуклеотидов, полученная по формуле (3). Для сравнения на этом рисунке показана оценка точности традиционно используемого метода распознавания сайтов YY-1 с помощью одних только частот нуклеотидов. Можно видеть, что при любых величинах ошибки I рода (недопредсказание) метод усреднения частот имеет меньшие ошибки II рода (перепредсказание), чем традиционный метод частот нуклеотидов. Это означает, что дополнительный учет частот динуклеотидов систематически увеличивает точность распознавания сайтов по сравнению с учетом одних только частот нуклеотидов.

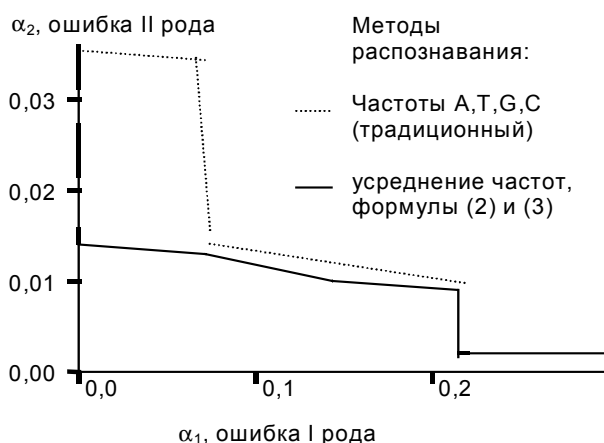


Рис. 3. Точность распознавания сайтов YY-1 по частотам олигонуклеотидов (сплошная линия) при любых ошибках I рода (недопредсказание) имеет меньшие ошибки II рода (перепредсказание), чем по частотам нуклеотидов (традиционный метод).

Базы знаний MATRIX и CONSENSUS содержат программы для распознавания сайтов с помощью консенсусов и весовых матриц. Эти программы применяются непосредственно для аннотации геномной ДНК (рис. 1). Применение программы распознавания сайтов связывания транскрипционного фактора YY-1 для аннотации интрона 6 гена *TDO2* человека описано в работе Меркуловой и др. [5].

Выборки функциональных сайтов ДНК и РНК, собранные в базе знаний SAMPLES, используются также для поиска значимых конформационных и физико-химических свойств В-ДНК сайтов. Для этого используется база данных PROPERTY, в текущей версии которой документировано 38 таких свойств ДНК. Примеры конформационных и физико-химических свойств В-ДНК из базы данных PROPERTY представлены в таблице 2.

Таблица 2  
Примеры конформационных и физико-химических свойств В-спирали ДНК

документ*	База данных PROPERTY свойство ДНК (обозначение, комментарий)	Шкала измерения		Лит. ссылка
		единицы	диапазон	
P0000003	Угол изгиба ДНК (Bend, теория)	Градус	4,62 ÷ 6,40	[6]
P0000005	Ширина малой бороздки (d, теория)	Ангстрем	4,62 ÷ 6,40	[6]
P0000014	Угол закрученности ДНК ( $\Omega$ , название "Twist")	Градус	27,7 ÷ 40,0	[7]
P0000020	Угол наклона ДНК ( $\delta$ , название "Direction")	Градус	-154 ÷ 180	[8]
P0000022	Температура плавления ( $T_m$ , теория)	°C	36,7 ÷ 136,1	[9]
P0000023	Частота контакта с кором нуклеосомы ( $P_N$ )	%	1,1 ÷ 18,4	[10]
P0000030	Угол пропеллер ( $\omega$ , свободная ДНК)	Градус	-17,3 ÷ -6,7	[11]
P0000038	Энтальпия ( $\Delta H$ , теория)	Kcal/mol	-11,8 ÷ -5,6	[12]

\* Для формулы (4): k – номер свойства ДНК, где «P00000k» – код документа базы данных.

Последовательность  $S=\{s_1...s_a...s_i s_{i+1}...s_b...s_L\}$  длины  $L$  характеризуется средним значениям  $X_{kab}(S)$  свойства  $X_k(s_i s_{i+1})$ , динуклеотидов  $s_i s_{i+1}$  на участке  $[a; b]$ :

$$X_{kab}(S)=\sum_{i=a,b-1} X_k(s_i s_{i+1})/(b-a). \quad (4)$$

здесь:  $1 \leq a \leq b \leq L$ ;  $1 \leq k \leq 38$  – номер свойства в базе данных PROPERTY (см. табл. 2).

Значимость  $k$ -го свойства В-спирали ДНК, усредненного на участке  $[a; b]$ ,  $X_{kab}$ , оценивается по степени различия между контрастными распределениями этих значений для  $N$  сайтов  $\{X_{kab}(S_n)\}_{n=1,N}$  и 1000 случайных последовательностей ДНК  $\{X_{kab}(Rand_m)\}_{m=1,1000}$ .

Идея используемого для этого метода иллюстрируется на рисунке 4 на примере ТАТА-бокса эукариот. На рисунке показаны выявленные нами контрастные распределения  $\{X_{kab}(S_n)\}$  и  $\{X_{kab}(Rand_m)\}$  для конформационного свойства «Угол изгиба ДНК, Bend» и физико-химического свойства «Температура плавления ДНК». Можно видеть, что изгиб ДНК (рис. 4,б) и необходимая для его образования низкая динамическая жесткость ДНК (температура плавления) (рис. 4,в) согласуются со строением ТВР/ТАТА-комплекса (рис. 4,а). Кроме того, важность для ТВР/ТАТА-комплекса как статического изгиба ДНК, так и ее динамической гибкости (низкая температура плавления) согласуются с результатами моделирования молекулярной динамики двойной спирали ДНК, содержащей ТАТА-бокс [14].

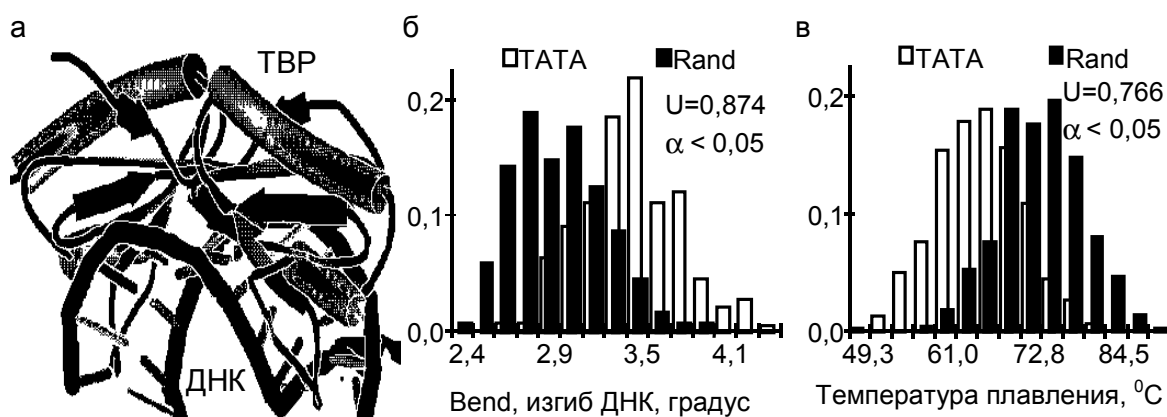


Рис. 4. Пространственная структура ТВР/ТАТА-комплекса [13] (а) и контрастные распределения  $\{X_{kab}(S_n)\}$  (светлые столбики) и  $\{X_{kab}(Rand_m)\}$  (темные столбики) для свойств ДНК (б) «Угол изгиба ДНК, Bend» и (в) «Температура плавления ДНК».

Поиск значимых конформационных и физико-химических свойств ДНК сайтов осуществляется следующим образом. Для распределений  $\{X_{kab}(S_n)\}$  и  $\{X_{kab}(Rand_m)\}$  с помощью 6 критериев проверяется значимость различия их средних, дисперсий, а также нормальность распределений. Каждый критерий проверяется 100 раз на различных 50%-подвыборках из  $\{X_{kab}(S_n)\}$  и  $\{X_{kab}(Rand_m)\}$ . Получается 600 оценок значимости  $\alpha_{pq}$   $1 \leq p \leq 6$  критериев в  $1 \leq q \leq 100$  испытаниях. Значимые оценки  $\alpha_{pq} < 0,05$  приписывают  $X_{kab}$  положительную оценку  $0 < u_{pq}(X_{kab}) \leq 1$ , незначимые –  $1 \leq u_{pq}(X_{kab}) \leq 0$ . В теории принятия решений [15]  $u_{pq}(X_{kab})$  называется «частной полезностью». Интегральная полезность равна ее средней частной оценке:

$$U(X_{kab})=\sum_{p=1,6} \sum_{q=1,100} u_{pq}(X_{kab})/600, \quad (5)$$

которая обладает двумя важными прогностическими свойствами:

$$U(X_{kab}) < 0 \quad \Leftrightarrow \quad X_{kab} \text{ НЕ СЛЕДУЕТ использовать для распознавания сайта; } \quad (6)$$

$$U(X_{kab}) > U(X_{qcd}) \geq 0 \Leftrightarrow \text{распознавать сайт ЛУЧШЕ с помощью } X_{kab}, \text{ чем } X_{qcd}. \quad (7)$$

Используя свойства (6) и (7), для всех возможных параметров  $\{k, a, b\}$  формулы (4) оцениваются полезности  $U(X_{kab})$  и находятся  $M$  независимых конформационных и физико-химических свойств  $\{X_m = X_{k(m); a(m); b(m)}\}$  с наибольшими  $U(X_m) > 0$ .

Аналогично рассмотренному выше методу распознавания сайтов с помощью усреднения частот олигонуклеотидов (формула (3)), строится метод распознавания сайтов по конформационным и физико-химическим свойствам ДНК:

$$D_{B\text{-ДНК}}(S) = \{\sum_{m=1, M} [X_m(S) - \alpha_m] / \beta_m\} / M, \quad (8)$$

где:  $\alpha_m = [\text{Mean}_{\text{SITE}}(X_m) + \text{Mean}_{\text{RANDOM}}(X_m)] / 2$ , и  $\beta_i = [\text{Mean}_{\text{SITE}}(X_m) - \text{Mean}_{\text{RANDOM}}(X_m)] / 2$  – нормирующие коэффициенты. При этом пикам положительных значений функции  $D_{B\text{-ДНК}}(S)$  над позицией  $i$  произвольной последовательности ДНК  $S$  указывает на локализацию потенциального функционального сайта в этой позиции.

Выявленные для функциональных сайтов значимые конформационные и физико-химические свойства ДНК, величины полезности этих свойств и программы распознавания сайтов по значимым свойствам их ДНК (формула (8)) документируются в базе знаний FEATURE. На рисунке 5 показаны пример документа базы знаний FEATURE, описывающий значимые конформационные и физико-химические свойства ДНК TATA-бокса, программа распознавания TATA-боксов с помощью этих свойств и результат работы этой программы для последовательности промотора гена *CYC1* изо-1-цитохрома С дрожжей (EMBL: X03472).

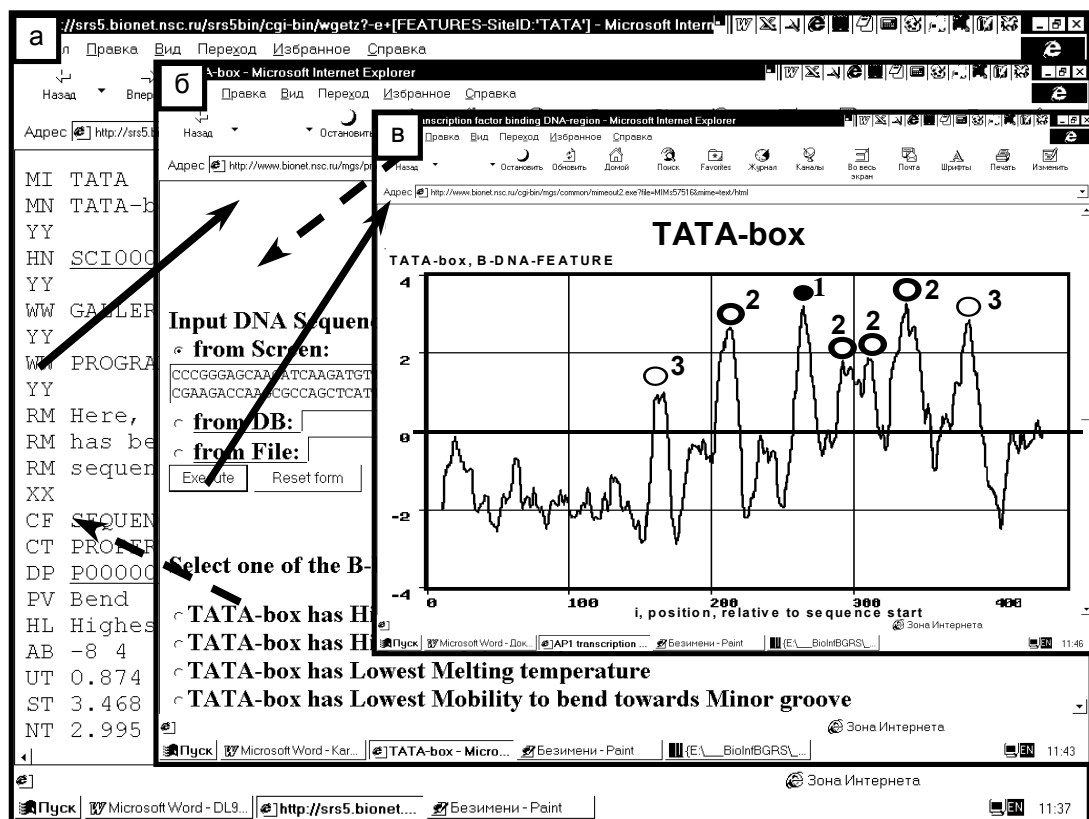


Рис. 5. База знаний FEATURES по конформационным и физико-химическим свойствам сайтов ДНК: а – документ с описанием свойств TATA-бокса; б – программа распознавания TATA-боксов по этим свойствам; в – результат работы этой программы для промотора гена *CYC1* изо-1-цитохрома С дрожжей (EMBL: X03472). Кружки: 1 – известный TATA-бок; 2 – документированные TATA-подобные боксы; 3 – ошибки II рода (перепредсказание). Пунктир – 95%-граница сходства.

В качестве сравнения с ранее предложенным эвристическим подходом к учету конформационных свойств В-спирали ДНК при распознавании сайтов связывания белков [6] отметим, что для указанного промотора этот подход предсказал 28 «ошибочных ТАТА-боксов» [6], тогда как наш подход, основанный на выявлении знаний путем анализа экспериментальных данных, дал лишь 2 ошибочных предсказания (рис. 5,в). Так, выявленные из экспериментальных данных знания строения В-спирали ДНК сайтов связывания белков определенного типа позволяют увеличить точность распознавания таких сайтов по сравнению с эвристическим учетом самых общих принципов ДНК-белковых взаимодействий.

Сравнительный анализ значимых конформационных и физико-химических свойств функциональных сайтов ДНК, накопленных в базе знаний FEATURES, позволяет выявлять ранее неизвестные закономерности ДНК-белковых взаимодействий.

В таблице 3 приведены все сайты связывания транскрипционных факторов, для которых значимым конформационным свойством ДНК является «угол Твист в комплексах ДНК-белок». Можно видеть, что два типа ДНК-связывающих доменов белков различаются по этому свойству сайтов их связывания на ДНК. По сравнению со случайными последовательностями ДНК сайты связывания факторов «Координационный цинк» имеют повышенный угол «Твист», сайты связывания факторов «Спираль-поворот-спираль» – пониженный. Согласно критерию Фишера, такое различие между транскрипционными факторами типов «Координационный цинк» и «Спираль-поворот-спираль» по углу «Твист» сайтов их связывания с ДНК является достоверным,  $\alpha < 0,05$ . Это означает, что сайты связывания разных транскрипционных факторов, содержащих конформационно сходные ДНК-связывающие домены, могут обладать сходными конформационными свойствами ДНК. Действительно, это – ранее не известная информация о природе и механизмах ДНК-белковых взаимодействий.

### 3.2. База знаний для предсказания величин активности сайтов ДНК и РНК

Регуляция экспрессии генов осуществляется благодаря связыванию регуляторных белков с сайтами на ДНК. Биологическая активность ДНК-белкового комплекса зависит от последовательности сайта, как это показано на рисунке 6,а для измеренных величин сродства белка ТВР к ДНК [16]. В базе данных ACTIVITY [17] собрано более 500 выборок последовательностей сайтов с известными величинами активности, аналогичных выборке, представленной на рисунке 6,а. Эти выборки анализируются с помощью специальной модификации описанного выше метода поиска значимых конформационных и физико-химических свойств ДНК сайтов, в котором вместо 6 критериев дискриминантного анализа используется 11 критериев корреляционного анализа.

В обозначениях формул (4) – (8) результатом анализа выборки  $\{S_n, F_n\}_{n=1, N}$  N сайтов S с активностью F являются M независимых свойств ДНК  $\{X_m = X_{k(m); a(m); b(m)}\}_{m=1, M}$  с наибольшими полезностями  $U(X_m, F) > 0$  предсказания этой активности. На рисунке 6 показаны примеры найденных корреляций конформационных свойств ДНК и величинами сродства белков ТВР, Cro и USF к сайтам их связывания на ДНК. С использованием линейно-аддитивной модели на основе выявленных значимых свойств ДНК  $\{X_m\}$  сайта, аналогично формулам (2) и (8), строится метод предсказания активности этого сайта по последовательности:

$$F(S) = F_0 + \sum_{m=1, M} \{\alpha_m \times (X_m(S))\}, \quad (9)$$

здесь:  $F_0$  и  $\alpha_m$  – коэффициенты множественной линейной регрессии.

Таблица 3

Транскрипционные факторы, для сайтов связывания которых значимым конформационным свойством ДНК является «угол Твист в комплексах ДНК-белок»

Транскрипционный фактор		Район [a; b]	Полез- ность, U	Среднее±ст. отк, градус		Значимость	
на- зва- ние	ДНК-связывающий домен			сайт	случайные последова- тельности	$\chi^2$	$\alpha$
NF-E2	Основной домен	-18; 8	0.756	33.9±0.2	34.1±0.2	187.8	0.005
USF	Основной домен	-13; 8	0.695	33.9±0.2	34.1±0.2	32.0	0.050
RF-X	Основной домен	-9; 16	0.586	34.0±0.2	34.1±0.2	123.9	0.005
CREB	Основной домен	-11; 6	0.462	34.0±0.2	34.1±0.3	34.7	0.050
AP-1	Основной домен	-10; 13	0.453	30.0±0.2	34.1±0.2	231.2	0.005
T3R	Координационный цинк	-14; 10	0.736	33.9±0.2	34.1±0.2	54.6	0.005
Sp-1	Координационный цинк	-14; 11	0.642	33.9±0.2	34.1±0.2	396.0	0.005
COUP	Координационный цинк	-11; 5	0.550	33.9±0.3	34.1±0.3	53.4	0.005
ER	Координационный цинк	-12; 14	0.536	33.9±0.2	34.1±0.2	43.9	0.050
RAR	Координационный цинк	-14; 12	0.448	33.9±0.2	34.1±0.2	28.8	0.050
GR	Координационный цинк	-9; 13	0.439	34.0±0.2	34.1±0.2	68.0	0.005
NF-κB	β-Платформа	-16; 0	0.452	33.9±0.2	34.1±0.3	116.1	0.005
E2	β-Платформа	-7; 5	0.311	33.9±0.3	34.1±0.3	34.4	0.050
СЕВРР	Основной домен	-9; 3	0.462	34.3±0.3	34.1±0.3	63.4	0.005
CP-1	Основной домен	-14; 12	0.504	34.2±0.1	34.1±0.2	73.0	0.005
E2F	Основной домен	-4; 14	0.765	34.5±0.2	34.1±0.3	92.2	0.005
IRF-1	Спираль-поворот-спираль	-14; 12	0.611	34.2±0.3	34.1±0.2	31.0	0.050
OCT	Спираль-поворот-спираль	-9; 4	0.621	34.3±0.3	34.1±0.3	154.6	0.005
EN	Спираль-поворот-спираль	-19; 0	0.684	34.3±0.2	34.1±0.2	26.7	0.050
HNF1	Спираль-поворот-спираль	-6; 10	0.725	34.5±0.3	34.1±0.3	95.8	0.005
HNF3	Спираль-поворот-спираль	-10; 6	0.748	34.4±0.3	34.1±0.3	138.2	0.005
TBP	β-Платформа	-4; 13	0.605	34.3±0.3	34.1±0.3	377.1	0.005
MEF-2	β-Платформа	-10; 2	0.904	34.4±0.4	34.1±0.3	81.0	0.005

Критерий  $\chi^2$  отличия между двумя распределениями, число степеней свободы  $\nu=15$ .

На рисунке 6,г показан результат предсказания величин ТВР/ДНК-средства, полученных на независимых данных (контроль) с помощью метода, основанного на линейной корреляции этого средства с шириной малой бороздки ДНК. Можно видеть, что этот результат является достоверным ( $\square < 0,01$ ). Кроме того, корреляция величин ТВР/ДНК-средства с шириной малой бороздки ДНК (рис. 6,б) согласуется с известной пространственной структурой ТВР/ДНК-комплекса (рис. 4,а), в котором белок ТВР взаимодействует с малой бороздкой В-спирали ДНК [13]. Применение программы предсказания величин ТВР/ДНК-средства к анализу природных, мутантных и синтетических ТАТА-боксов описано в работе Савинковой и др. [16].

Выявленные корреляции конформационных и физико-химических свойств ДНК с активностью сайта и программы предсказания этой активности документируются в базе знаний KNOWLEDGE. На рисунке 7 показано применение знания-программы предсказания USF/ДНК-средства к анализу промотора гена HMG ко-редуктазы хомяка.

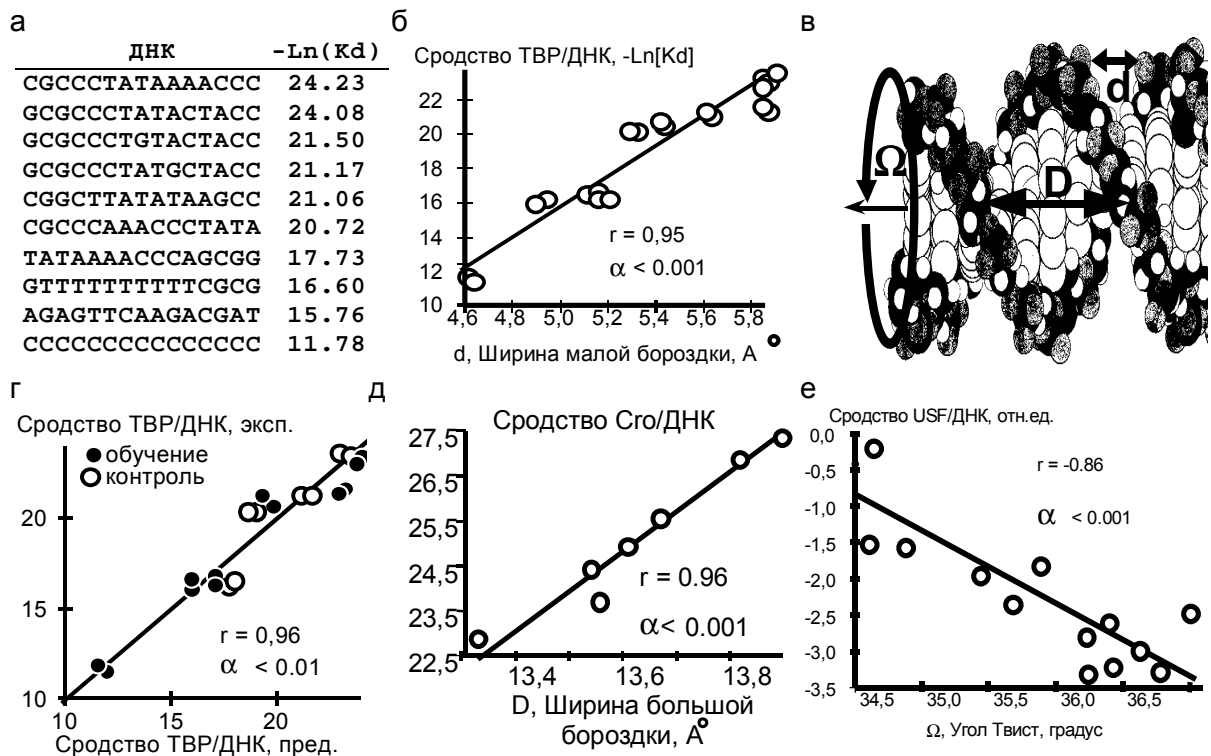


Рис. 6. Экспериментальные данные [16] о сродстве ТВР-белка к ДНК (а) и выявленная на их основе (б) корреляция ТВР/ДНК-сродства с шириной малой бороздки ДНК (в). Основанный на этой корреляции метод (г) достоверно предсказал ТВР/ДНК-сродство. Корреляции сродства белков (д) Cro и (е) USF к ДНК и свойств конформации ДНК (в).

### 3.3. Пример использования баз знаний для комплексного анализа промоторов

Предлагаемые базы знаний по молекулярной биологии созданы, прежде всего, для накопления множества компьютерных программ анализа ДНК. Поэтому представляется интересным вопрос: что нового может дать комплексный анализ геномной ДНК с помощью нескольких качественно разных программ? На рисунке 8 показано сопоставление результатов анализа 194 последовательностей ТАТА-содержащих промоторов с помощью программ распознавания старта транскрипции, нуклеосомной ДНК и ТАТА-бокса, а также с помощью программы предсказания ТВР/ДНК-сродства. Можно видеть, что сходство ДНК этих промоторов со стартами транскрипции имеет сложный профиль (рис. 8.а). На участке от начала корового промотора в позиции -70 относительно старта транскрипции до ТАТА-бокса в позиции -30 этот профиль является возрастающим трендом в направлении старта транскрипции (позиция +1, коэффициент линейной корреляции  $r=0,966$ ). Этот тренд обрывается отрицательным пиком над так называемым ТАТА/TSS-спейсером, разделяющим ключевые сигналы транскрипции, ТАТА-бокс и старт транскрипции.

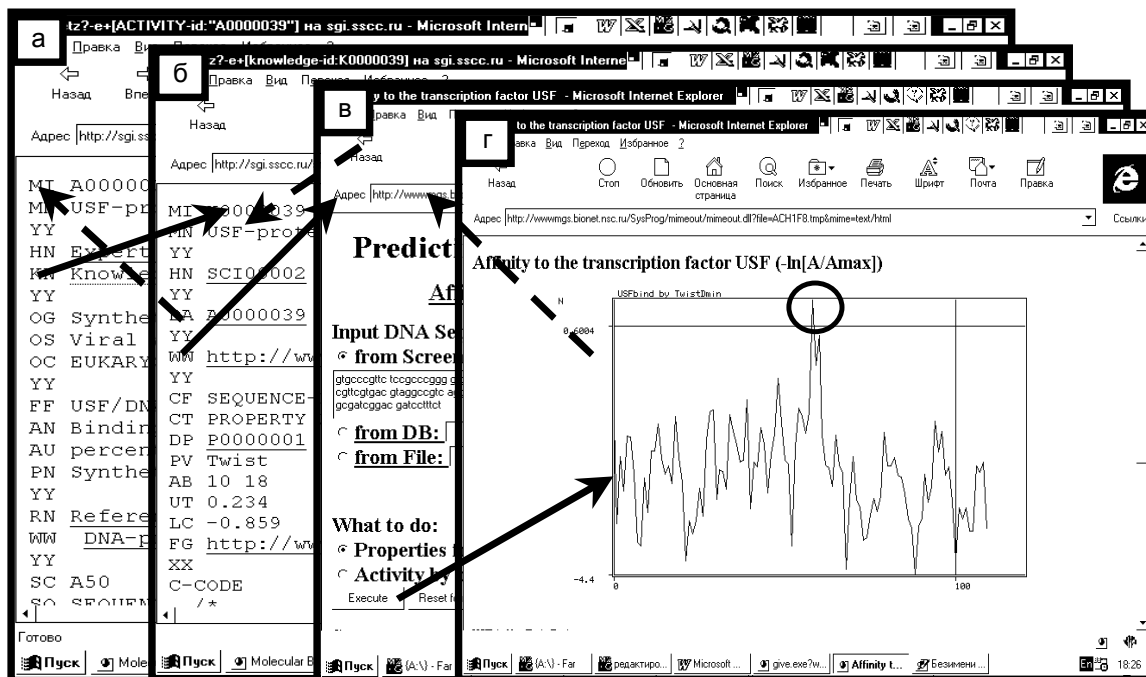


Рис. 7. База знаний KNOWLEDGE по корреляциям свойств ДНК с активностью сайтов: **а** – документ базы данных ACTIVITY, USF/ДНК-средство; **б** – документ базы знаний KNOWLEDGE, корреляции этого средства со свойствами ДНК; **в** – программа предсказания USF/ДНК-средства; **г** – результат работы этой программы для промотора гена HMG ко-редуктазы хомяка (EMBL: M15960). Кругок – экспериментально известный сайт USF.

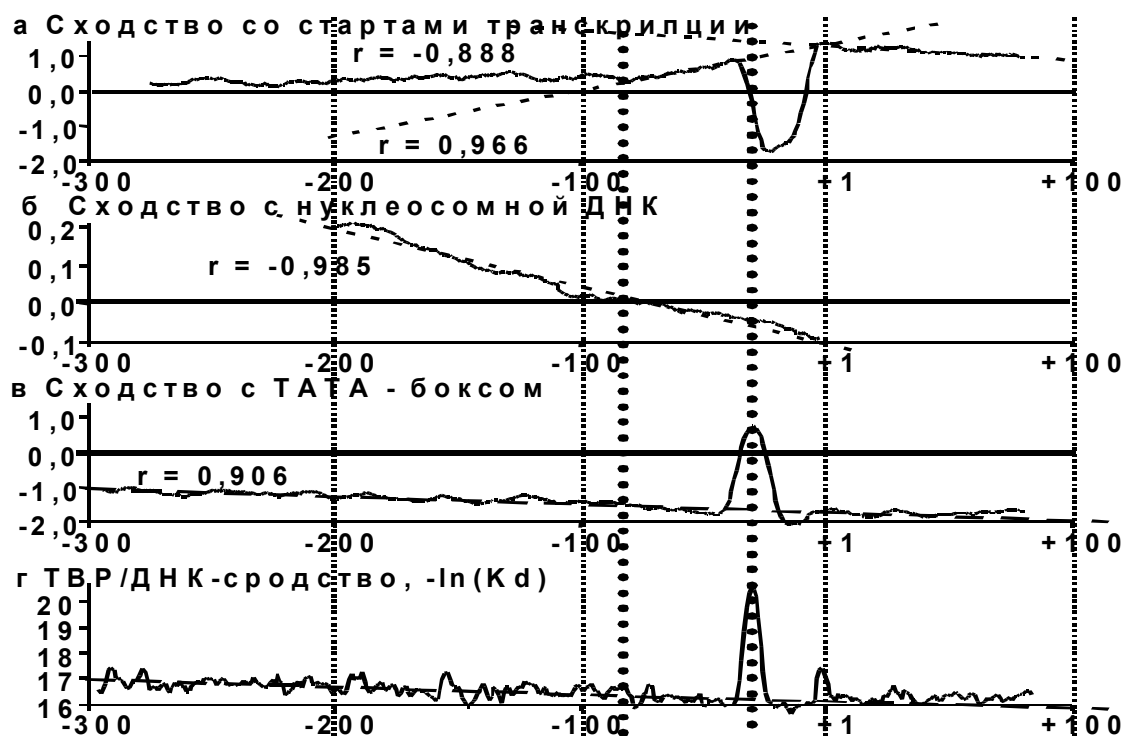


Рис. 8. Сопоставление результатов анализа «ТАТА+»-промоторов с помощью программ распознавания (**а**) старта транскрипции, (**б**) нуклеосомной ДНК, (**в**) ТАТА-бокса, а также с помощью (**г**) программы предсказания ТВР/ДНК-средства.

На рисунке 8,б для этих же промоторов показан профиль их сходства с нуклеосомной ДНК, который является убывающим трендом ( $r=0,985$ ), пересекающим линию «нулевого» сходства в начале корового промотора. Это означает, что район корового промотора, имеющий отрицательное сходство с нуклеосомной ДНК, является менее предпочтительным для формирования нуклеосомы по сравнению с предшествующим ему участком дистального промотора.

Наконец, профили сходства промоторов с ТАТА-боксом (рис. 8,в) и сродства ДНК к белку ТВР (рис. 8,г) имеют пики над оптимальной для ТАТА-боксов позицией -30. Таким образом, несмотря на различия между «сходством с сайтом» (формула (8)) и «активностью сайта» (формула (9)), обе эти значимые особенности ТАТА-боксов согласуются как одна с другой, так и с экспериментальными данными. Можно видеть, что результаты распознавания различных регуляторных элементов (проксимальный промотор, коровый промотор, старт транскрипции, начало 1-го экзона) определенной группы генов показывают сложную картину инициации транскрипции. Таким образом, базы знаний, накапливающие знания-программы для распознавания регуляторных элементов, открывают возможности комплексного анализа геномной ДНК, которые отсутствуют при более простом традиционном анализе профилей конформационных свойств В-ДНК вблизи функциональных сайтов [18].

Результаты работы показывают, что информационно-поисковая система [базы данных]  $\Leftrightarrow$  [базы знаний]  $\Leftrightarrow$  [программы] позволяет для новых экспериментальных данных найти сходные данные, для которых были выявлены закономерности и программы проверки этих закономерностей. Сопоставление результатов проверки нескольких закономерностей дает картину исследуемого явления.

Работа была поддержана грантами РФФИ 98-07-90126 и 98-07-91078.

### Список литературы

1. Kolchanov N.A., Ponomarenko M.P., Frolov A.S. et al. Integrated databases and computer systems for studying the eukaryotic gene expression // *Bioinformatics*. 1999. V. 15, № 7/8. P. 669-686.
2. Kolchanov N.A., Ananko E.A., Podkolodnaya O.A. et al. Transcription Regulatory Regions Database (TRRD): its status in 1999 // *Nucleic Acids Res.* 1999. V. 27, № 1. P. 303-306.
3. Lawrence C. Toward the unification of sequence and structural data for identification of structural and functional constraints // *Comput. Chem.* 1994. V. 18. P. 255-258.
4. Heinemeyer T., Chen X., Karas H. et al. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms // *Nucleic Acids Res.* 1999. V. 27, № 1. P. 318-322.
5. Васильев Г.В., Меркулов В.М., Горшкова Е.В. и др. Точковые мутации в интроне 6 гена триптофаноксигеназы, ассоциированные с рядом психических расстройств, приводят к изменению спектра ядерных белков, связывающихся с этим районом // Настоящий сборник.
6. Karas H., Knuppel R., Schulz W. et al. Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements // *Comput. Applic. Biosci.* 1996. V. 12. P. 441-446.
7. Suzuki M., Amano N., Kakinuma J., Tateno M. Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA // *J. Mol. Biol.* 1997. V. 274. P. 421-435.
8. Shpigelman E.S., Trifonov E.N., Bolshoy A. CURVATURE: software for the analysis of curved DNA // *Comput. Appl. Biosci.* 1993. V. 9. P. 435-440.
9. Gotoh O., Tagashira Y. Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles // *Biopolymers*. 1981. V. 20. P. 1033-1042.
10. Satchwell S.C., Travers A.A. Asymmetry and polarity of nucleosomes in chicken erythrocyte chromatin // *EMBO J.* 1989. V. 8, № 1. P. 229-238.

11. Gorin A.A., Zhurkin V.B., Olson W.K. B-DNA twisting correlates with base-pair morphology // J. Mol. Biol. 1995. V. 247. P. 34–48.
12. Sugimoto N., Nakano S., Yoneyama M., Honda K. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes // Nucleic Acids Res. 1996. V. 24. P. 4501–4505.
13. Juo Z.S., Chiu T.K., Leiberman P.M. et al. How proteins recognize the TATA box // J. Mol. Biol. 1996. V. 261. P. 239–254.
14. Flatters D., Lavery R. Sequence-dependent dynamics of TATA-Box binding sites // J. Biophys. 1998. V. 75. P. 372–381.
15. Fishburn P.C. Utility theory for decision making. NY: John Wiley and Sons, 1970.
16. Савинкова Л.К., Соколенко А.А., Пау В.А. и др. Структурно-функциональная гомология РНК-полимераз прокариот и эукариот // Настоящ. сборник. С.
17. Ponomarenko M.P., Ponomarenko J.V., Frolov A.S. et al. Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins // Bioinformatics. 1999. V. 15, № 7/8. P. 687–703.
18. Baldi P., Chauvin Y., Brunak S. et al. Computational applications of DNA structural scales // Proc. of the 6th Intern. Conf. on intelligent systems for molecular biology (ISMB-98). 1998. Montreal: AAAI Press. P. 35–42.