

# Информационная система FrameProt по пространственным структурам ДНК, РНК и белков в составе GeneExpress

В.А. Иванисенко, Д.А. Григорович, Д.А. Афонников, Д.А. Куропатов, В.П. Валуев, Н.А. Колчанов.  
Институт Цитологии и Генетики СО РАН, Лаборатория Теоретической Генетики, Новосибирск.  
E-mail: salix@bionet.nsc.ru, odip@bionet.nsc.ru

## 1 Введение

Разработанная в лаборатории Теоретической Генетики ИЦиГ СО РАН компьютерная система GeneExpress [18] обеспечивает интеграцию нескольких десятков баз данных, баз знаний и сотен программ для обработки информации по регуляции экспрессии генов. В число крупных информационно-программных модулей GeneExpress, в частности, входят база данных по регуляции транскрипции генов эукариот TRRD, системы по распознаванию и предсказанию активности сайтов B-DNA Video и Activity, система, описывающая генные и метаболические сети [9, 15]. Существующий в настоящее время вариант GeneExpress описывает верхние ярусы системы молекулярно-генетических процессов, имеющих ключевое значение при функционировании молекулярно-генетических систем [2].

Базовыми элементами во всех молекулярно-генетических системах живых организмов являются ДНК, РНК и белки. ДНК служит материальным носителем наследственной информации, в процессе реализации которой производится синтез молекулы мРНК, в свою очередь служащей матрицей для дальнейшего синтеза полипептидной цепи, спонтанно укладываемой в пространственную структуру белка. Белки за счет своей уникальной пространственной структуры обеспечивают основные процессы метаболизма (т. е. биосинтеза и деградации всех молекулярных компонент организмов). Именно белки в составе различных макромолекулярных комплексов, таких как РНК-полимеразы, сплайсомеры и рибосомы обеспечивают считывание информации от гена к белковой молекуле. В свою очередь, молекулы ДНК и РНК для своего функционирования также должны иметь уникальную пространственную структуру.

Целью настоящей работы является развитие нового информационного ресурса FrameProt по пространственным структурам ДНК, РНК и белков и их структурным и функциональным особенностям в составе системы GeneExpress.

Знание информации о пространственной структуре ДНК, РНК и белков необходимо для решения широкого круга задач геномной и белковой инженерии, биотехноло-

гии, искусственного синтеза молекулярно-генетических систем, а также для различных медико-биологических приложений.

## 2 Основные направления работ

Основными направлениями работ при разработке FrameProt были:

- создание базы данных пространственных структур;
- дочерних баз данных;
- программ анализа пространственных структур;
- создание системы конвертации из базы PDB в новую базу ENPDB;
- обеспечение программами визуализации молекул;
- интеграция всех ресурсов в единую систему;
- обеспечение доступа к системе через Интернет.

На рис. 1 приведена схема пилотной версии компьютерной системы FrameProt по пространственным структурам ДНК, РНК и белков. Система состоит из связанных между собой блоков. Ядром системы является база данных пространственных структур EnPDB [19].

### 2.1 Ядро системы FrameProt — база данных пространственных структур

Информация о пространственной структуре ДНК, РНК и белков накапливается в Protein Data Bank [6]. Этот банк данных является единственным в мире официальным источником научной информации об известных пространственных структурах макромолекул. В PDB содержится информация об аминокислотной последовательности белка, называемой первичной структурой (это последовательность символов), информация о вторичной структуре белка (локальной укладке полипептидной цепи в пространстве в виде альфа-спиралей и бета-нитей), а также информация о координатах атомов, образующих этот белок. Типичный белок содержит до нескольких тысяч атомов, и для каждого из них приводятся координаты X, Y, Z. Доступ к этой информации для некоммерческих организаций, проводящих фундаментальные исследования, свободный и может быть осуществлен через Интернет. Однако в период создания PDB число расшифрованных структур было невелико. В связи с этим в то

Первая Всероссийская научная конференция  
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:  
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,  
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ  
19 - 21 октября 1999 г., Санкт-Петербург

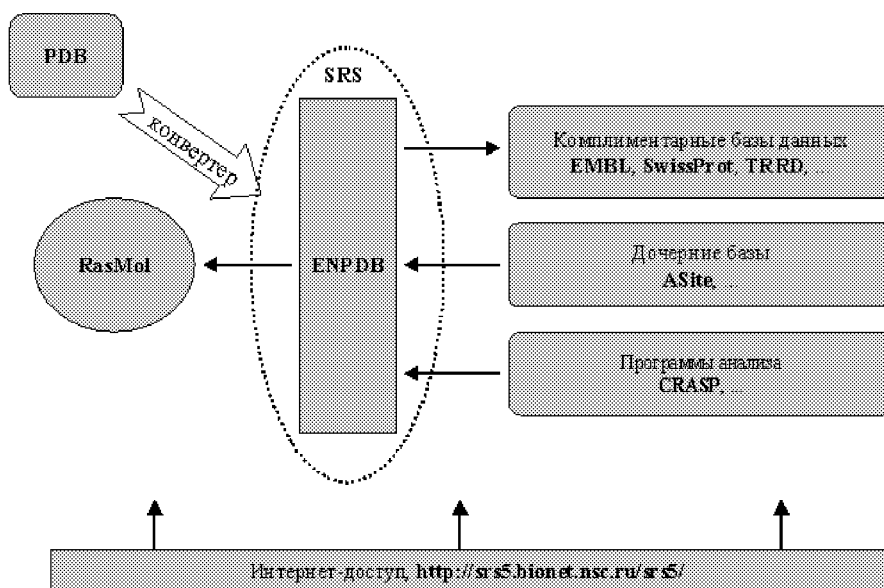


Рис. 1: Схема компьютерной системы FrameProt по пространственным структурам ДНК, РНК и белков. Стрелками обозначены связи между базами данных, а также программы.

время не стояла задача автоматизированного поиска и доступа к информации по пространственным структурам. Разработчики PDB сделали акцент на полноту представления структурных данных, методов и условий экспериментов, при которых эти данные получены, а также функциональных особенностей макромолекул. Структура же самой базы была выбрана наиболее оптимальной для первичного накопления данных с учетом возможностей добавления новой информации. Например, информация для разных макромолекул в PDB разнесена по отдельным файлам, кроме того, существует ряд зарезервированных полей, которые, как предполагалось, обеспечат “эволюцию” базы вслед за эволюцией экспериментальных и теоретических технологий извлечения новых знаний.

За последние несколько лет произошел качественный скачок техники синтеза, выделения и кристаллизации биологических макромолекул. В результате чего увеличилось число расшифрованных структур. В настоящее время число таких структур достигло более 10000 и продолжает расти, увеличиваясь каждый год более чем в два раза. Обеспечение эффективного поиска в таких огромных объемах данных возможно только при использовании современных компьютерных технологий. Одной из таких технологий для поиска и доступа к молекулярно-биологическим базам данных является SRS. Эффективное использование SRS напрямую зависит от степени формализации исходных данных. Поэтому на первом этапе ставилась задача трансформации базы PDB в базу EnPDB, имеющую более формализованную форму. Данная задача включала в себя следующие подзадачи:

- классификация информации, содержащейся в PDB;
- выделение значимой информации;
- построение алгоритма преобразования;
- реализация алгоритма в виде программы конвертации.

Кроме этого ставилась задача создания связей базы EnPDB с другими базами данных и задача добавления новой информации, расширяющей возможности поиска.

## 2.2 Программы визуализации и анализа

Одной из проблем в использовании информации о пространственной структуре макромолекулы заключается в том, что она не может быть осмыслена без вспомогательных программных средств. Наиболее известной среди свободно распространяемых средств визуализации пространственных структур является программа RasMol. С ее помощью пользователь может визуализировать отдельный белок на своем компьютере, взяв запись из базы PDB. Таким образом, следующей задачей осуществления проекта стояла задача интеграции EnPDB с программой визуализации RasMol.

Следует однако подчеркнуть, что при решении фундаментальных и практических задач молекулярной биологии требуется получение гораздо более детальной информации о пространственной структуре белков, чем просто ее отображение. Например, пользователя может интересовать вопрос о наличии полостей в пространственной структуре белков. Его может интересовать картина распределения положительно и отрицательно заряженных остатков на поверхности белков, наличие сходства между пространственными структурами различных белков, особенности локального окружения активных центров определенного типа. Информация об этих и многих десятках, если не сотнях, других исключительно важных особенностей пространственной структуры белков в базе PDB отсутствует. В связи с этим, ставилась задача разработки программ анализа пространственных структур и их интеграции с базой EnPDB

## 2.3 Дочерние базы данных

Следует еще раз особо подчеркнуть огромные объемы информации, уже накопленной в PDB. В настоящее время суммарный объем файлов PDB составляет более 5 гигабайт. Ожидается, что при имеющемся темпе роста этот объем данных уже через 10 лет может достигнуть одного терабайта. При таких объемах информации расчет одной характеристики из числа перечисленных ранее для всех белков из PDB является исключительно затратной задачей, требующей многих часов, дней, или даже недель вычислений на быстродействующих компьютерах. Поэтому, еще одной важной задачей, которая решалась нами в рамках настоящего проекта, являлось создание семейства дочерних баз данных и их интеграция с EnPDB. Каждая из дочерних баз данных содержит информацию об отдельных особенностях белков, хранящихся в EnPDB.

## 2.4 Комплиментарные базы данных

Следует подчеркнуть, что любой белок реально работает во вполне конкретном молекулярно-генетическом окружении. Поэтому его функция может быть понята только в контексте той конкретной клеточной и организменной ситуации, которая характерна для его функционирования. Так, для понимания функционирования регуляторных белков, контролирующих транскрипцию генов, необходима информация о сайтах связывания с ДНК [13]. Например, функция ТВР-белка, играющего ключевую роль в процессах инициации транскрипции, может быть понята только при одновременном рассмотрении структуры этого белка и структуры соответствующего сайта связывания на ДНК (так называемого ТАТА-бокса). При изучении механизмов функционирования этого белка требуется огромное количество дополнительной информации, которая не содержится в базе данных PDB, в том числе информация об изгибной жесткости того участка ДНК, с которым этот белок взаимодействует, а также об аффинности различных ТАТА-боксов к ТВР-белку. При изучении молекулярных механизмов функционирования белков исключительно важной является информация о генных сетях, в составе которых эти белки функционируют. Очень полезную информацию о функции белков дают такие базы данных, как SwissProt и PIR [5]. Эти базы данных предназначены для накопления информации о первичной структуре белков и содержащихся в них функциональных сайтах. Не менее ценную дополнительную информацию о белке можно получить, исследуя первичную структуру кодирующего его гена, которая накапливается в базах данных EMBL и GeneBank. Наконец, для получения сведения об особенностях регуляции экспрессии гена, кодирующего определенный белок, исследователь может использовать базу данных TRRD, также разрабатываемую в лаборатории Теоретической генетики ИЦиГ СО РАН.

Таким образом, одновременная работа пользователя — молекулярного биолога с большим количеством баз данных, комплиментарных PDB, будет приводить к принципиально новому качеству получаемых результатов. Поэтому, следующей важной задачей, решаемой в рамках проекта, была интеграция базы данных EnPDB с комплиментарными ей базами данных, содержащими самую разнообразную дополнительную информацию о структурно-функциональной организации исследуемого белка и особенностях его функционирования в составе молекулярно-генетических систем клеток.

В настоящее время нами получены первые результаты в перечисленных выше направлениях. Использование SRS позволило нам быстро создать систему, включающую Интернет — доступный вариант базы EnPDB, снабженную средством визуализации пространственных структур, программой анализа функционально-скоррелированных замен в белках и интегрированную с большим количеством комплиментарных баз данных. Подробное описание о каждом элементе системы FrameProt и перспективах ее развития, как в целом, так и каждого из ее ресурсов дано ниже.

## 2.5 Применение SRS при создании FrameProt

SRS (Sequence Retrieval System) был разработан в Европейской молекулярно-генетической лаборатории (EMBL), как комплекс программ для обеспечения индексирования, поиска, интеграции молекулярно-генетических баз данных, а также для универсального доступа к ним. В настоящее время более 100 основных молекулярно-генетических баз данных уже установлены под SRS. Доступ к системе SRS осуществляется через Web-интерфейс с помощью браузера, например Internet Explorer или Netscape Navigator.

В SRS используется объектно-ориентированный язык Icarus. При создании EnPDB на этом языке описывалась структура базы данных, а именно были описаны поля базы данных, их названия и тип (числовой, текстовый, идентификатор, дата). С помощью программы на языке Icarus было проведено индексирование данных и программировался внешний вид представления базы EnPDB для пользователя. Индексы были созданы для 40 различных полей базы данных. Средства индексирования данных позволяют осуществлять быстрый поиск. Исходные данные были расположены в одном текстовом файле, содержащем множество записей EnPDB. Гибкие средства SRS представления данных в виде HTML-файла позволили нам выполнить интеграцию базы EnPDB на уровне гиперссылок с другими комплиментарными базами данных и программами, включая программу визуализации пространственных структур.

Аналогичная работа с помощью SRS проводилась и с другими базами в составе системы FrameProt.

## 3 База данных EnPDB

Как уже отмечалось выше вся информация о пространственных структурах биологических макромолекул ДНК, РНК и белков концентрируется в базе PDB. Для каждой макромолекулы в PDB отводится одна запись, представленная в отдельном текстовом файле с именем, соответствующим ее идентификатору в PDB. Пример записи PDB приведен на рис. 2.

На данном рисунке приводится урезанная запись, поскольку оригинал, как практически каждая запись в PDB имеет гораздо больший размер по сравнению с допустимым объемом настоящей статьи. Каждая запись содержит широкий спектр характеристик экспериментов по биологическому и химическому синтезу макромолекулы, описание ее биологической функции на молекулярном, клеточном и организменном уровне, коллекцию координат атомов, входящих в состав макромолекулы и другие функциональные и структурные характеристики.

Информацию, накапливаемую в PDB можно классифицировать следующим образом:

```

HEADER      PROTEIN/DNA                      13-JAN-94  2BOP
JNHL       AUTH      R.S.HEDDE, S.R.CROSSMAN, L.A.LAMMINS, P.E.SICLER
JNHL       TITL      CRYSTAL STRUCTURE AT 1.7 ANGSTROMS OF THE BOVINE
JNHL       TITL 2    PAPILLOMAVIRUS-1 E2 DNA-BINDING DOMAIN BOUND TO
JNHL       TITL 3    ITS DNA TARGET
JNHL       REF       NATURE                      V. 359   505 1992
JNHL       REFN      ASTH NATUAS UK ISSN 0028-0836                      006
SQRES      1 A 85   SER CYS PHE ALA LEU ILE SER GLY THR ALA ASN GLN VAL
SQRES      2 A 85   LYS CYS TYR ARG PHE ARG VAL LYS LYS ASN HIS ARG HIS
SQRES      3 A 85   ARG TYR GLU ASN CYS THR THR THR TRP PHE THR VAL ALA
SQRES      4 A 85   ASP ASN GLY ALA GLU ARG GLN GLY GLN ALA GLN ILE LEU
SQRES      5 A 85   ILE THR PHE GLY SER PRO SER GLN ARG GLN ASP PHE LEU
SQRES      6 A 85   LYS HIS VAL PRO LEU PRO PRO GLY MET ASN ILE SER GLY
SQRES      7 A 85   PHE THR ALA SER LEU ASP PHE
SQRES      1 B 17   C C G A C C G A C G T C G
SQRES      2 B 17   G T C G
HELIX      1 A1 THR A 334 HIS A 349 1
HELIX      2 A2 SER A 382 VAL A 393 1
SHEET      B 4 TYR A 353 ASP A 365 0
SHEET      B 4 GLU A 369 GLY A 381 -1
SHEET      B 4 SER A 326 THR A 334 -1
SHEET      B 4 MET A 399 PHE A 410 -1
CRYST1    64.600 64.600 64.600 61.80 61.80 61.80 R 3 2 (RHOM) 6
ORIGX1    1.000000 0.000000 0.000000 0.000000 0.000000
ORIGX2    0.000000 1.000000 0.000000 0.000000 0.000000
ORIGX3    0.000000 0.000000 1.000000 0.000000 0.000000
SCALE1    0.015480 -0.008300 -0.005951 0.000000
SCALE2    0.000000 0.017555 -0.005951 0.000000
SCALE3    0.000000 0.000000 0.018546 0.000000
ATOM       1 N SER A 326 103.777 74.304 20.170 1.00 21.58 N
ATOM       2 CA SER A 326 102.613 74.991 20.586 1.00 18.59 C
ATOM       3 C SER A 326 101.631 74.211 21.431 1.00 17.75 C
ATOM       4 O SER A 326 101.653 74.549 22.634 1.00 18.51 O
ATOM       5 CB SER A 326 102.334 76.356 20.223 1.00 20.28 C
ATOM       6 OG SER A 326 101.401 76.656 19.086 1.00 24.53 O
ATOM       7 N CYS A 327 100.778 73.378 20.953 1.00 16.74 N
ATOM       8 CA CYS A 327 99.773 72.639 21.771 1.00 14.30 C
ATOM       9 C CYS A 327 100.306 71.215 22.046 1.00 12.49 C
ATOM      10 O CYS A 327 100.684 70.528 21.068 1.00 12.79 O
ATOM      21 CE1 PHE A 328 103.964 72.396 25.761 1.00 14.07 C
ATOM      22 CE2 PHE A 328 104.156 70.332 27.047 1.00 13.44 C
ATOM      23 CZ PHE A 328 104.395 71.733 26.894 1.00 13.72 C

```

Рис. 2: Пример одной записи в базе PDB.

- **Поисковая** — не имеющая биологического смысла, но однозначно определяющая один вход либо их некоторое число, например, идентификатор записи, дата создания записи и т. д.;
- **Поисково-описательная** — перечислимая, лаконично выраженная, имеющая биологический смысл, однозначно определяющая одну макромолекулу или группу макромолекул, объединенных общим функциональным либо структурным свойством, например, ключевые слова, способ определения третичной структуры и т. д.;
- **Описательная** — представленная в виде свободного описания особенностей макромолекул, а также координаты атомов. Прямое использование такой информации для поиска либо затруднительно, либо не имеет смысла, как в случае координат атомов.

При трансформации PDB в новую базу EnPDB [19] нами была отобрана информация, которая позволяет проводить полноценный поиск и устанавливать связи с другими базами, содержащими дополнительную информацию о макромолекуле. Список полей записи EnPDB как он выглядит в браузере приведен на рис. 3.

В EnPDB не была включена информация, относящаяся к разряду чисто описательной. Это позволило значительно уменьшить объем базы и сделать ее структуру легко адаптируемой для наиболее полной индексации средствами SRS. Для целей адаптации также была проведена декомпозиция некоторых полей PDB в ряд новых полей, содержащих однородную информацию. Например, поле HEADER было разбито на три поля: ID, HEADER и DATA (см. рис. 3.).

Name	Short Name	Type	No of Keys	No of Entry References	Indexing Date	Status
ID	id	ID	4280	4280	7/30/99	ok
Header	hdr	string	723	8494	7/30/99	ok
Date	dt	int	1009	4280	7/30/99	ok
Title	tit	string	5103	35052	7/30/99	ok
Compound	cpd	show				not indexed
Molecule	mol	string	2650	10672	7/30/99	ok
Synonym	snm	string	1442	3778	7/30/99	ok
EC	ec	string	388	1835	7/30/99	ok
BioUnit	bun	string	126	941	7/30/99	ok
Gene	gen	show				not indexed
MolSource	ms	string	3	4280	7/30/99	ok
Source	src	string	3281	21770	7/30/99	ok
Synthesis	snt	string	1785	13189	7/30/99	ok
Keyword	kw	string	2891	29035	7/30/99	ok
Technique	tch	string	22	9090	7/30/99	ok
Author	aut	string	4668	16316	7/30/99	ok
Id	id	show				not indexed
JrefAuthor	jau	string	8117	34509	7/30/99	ok
JrefTitle	jti	string	6962	90638	7/30/99	ok
JrefRef	jre	string	362	8885	7/30/99	ok
JrefVolume	jvo	string	391	8168	7/30/99	ok
JrefYear	jye	int	32	7362	7/30/99	ok
Remark 1	rm	show				not indexed
Resolution	res	real	174	3461	7/30/99	ok
ChemAmount	cha	int	22	4277	7/30/99	ok
ChemSize	chs	int	587	9061	7/30/99	ok
ProtAmount	pta	int	130	4280	7/30/99	ok
ProtSize	psa	int	117	4280	7/30/99	ok
DnaRnaAmount	dna	int	8	4277	7/30/99	ok
ProtFormAmount	pra	int	23	4277	7/30/99	ok
ProtAmount	pta	int	46	4280	7/30/99	ok
Heteroatom	htg	string	2317	15613	7/30/99	ok

Рис. 3: Список полей (индексов) в базе EnPDB. Для индексов приведено число различных значений и общее число значений. Так же приводится другая служебная информация.

Вместе с тем, добавлена была и новая информация. Значительно было расширено количество ссылок на комплементарные базы данных. Введены новые поля, характеризующие структурные особенности макромолекулы: количество альфа-спиралей, бета-нитей, полипептидных цепей, нуклеиновых кислот, гетероатомов и т. д. Добавление такой информации позволило провести интеграцию EnPDB с внутренними базами GeneExpress, на текущий момент времени это базы TRRD и ASPD, а также с рядом внешних баз данных (Swiss-Prot, Pir, Transfac, EMBL, GeneBank, и пр.). С использованием средств SRS реализована возможность перехода из EnPDB в эти базы данных с помощью гипертекстовых ссылок (см. рис. 4).

Кроме этого реализована возможность непосредственного вызова программы визуализации пространственных структур RasMol (см. рис. 5).

Индексация средствами SRS структурной информации обеспечила возможность проводить поиск с использованием критериев по структурным характеристикам и формировать выборки белков, нуклеиновых кислот или их комплексов, удовлетворяющих данным критериям. Например, работая с комбинацией полей ProteinAmount (количество белков), DnaRnaAmount (количество нуклеиновых кислот), HetAmount (количество гетероатомов) возможно создавать запросы на поиск всех записей, содержащих пространственные структуры: мономерных молекул белков, полисубъединичных белков, белок-белковых комплексов, комплексов белок-нуклеиновая кислота, комплексов белок-лиганд. Пример одного из таких запросов приведен на рис. 6. Формирование таких выбо-

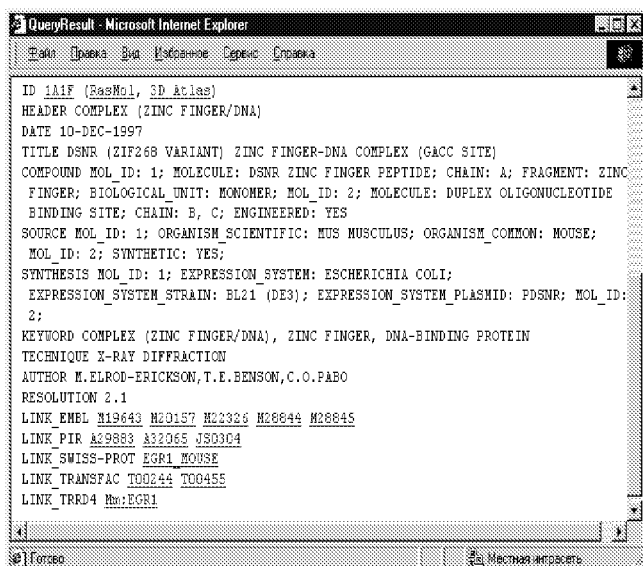


Рис. 4: Пример гиперссылок на внешние базы данных.

рок или им подобных является одной из наиболее трудоемких задач в структурно-функциональном анализе белков.

В общем виде, информацию, содержащуюся в EnPDB можно классифицировать по ее функциональной нагрузке следующим образом:

- Источник происхождения макромолекул и условия экспериментов по их синтезу;
- Эксперименты по определению пространственной структуры;
- Структурно-функциональная организация макромолекул;
- Информация, необходимая для интеграции с другими базами;
- Литературные ссылки.

Нами реализован также доступ к PDB несколькими путями. В карточке ENPDB в поле идентификатора имеется три гиперссылки. Одна из них ведет на соответствующий вход базы PDB, расположенной в системе SRS5. Вторая ссылка ведет к вызову программы визуализации RasMol. Третья ссылка ведет на соответствующий вход базы 3D-Atlas.

Таким образом обеспечивается совместимость с программами визуализации и анализа, которые работают с оригинальным форматом PDB. Кроме того, пользователь получает доступ к той информации, которая не была включена в EnPDB.

#### 4 Дочерняя база данных Asite

Разработанная нами база Asite является дочерней базой по отношению к базе пространственных структур биологических макромолекул. В ней накапливается уникальная информация по структуре активных сайтов в белках,

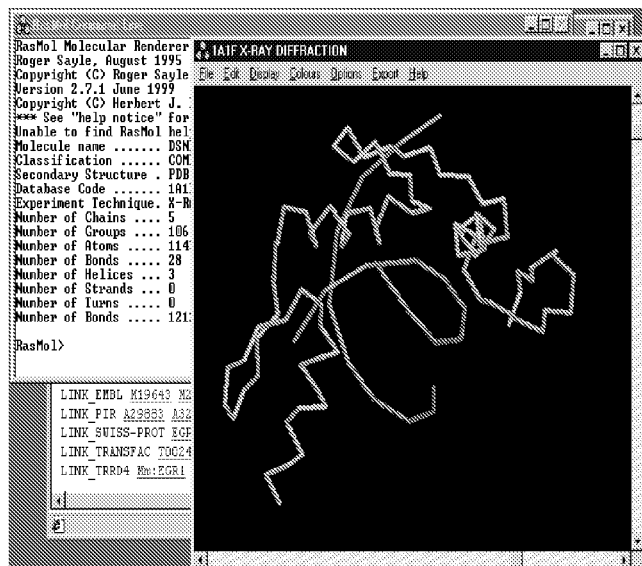


Рис. 5: Визуализация входа EnPDB с помощью программы RasMol.

не имеющая аналогов в существующих на сегодняшний момент информационных Интернет-доступных ресурсах. При разработке дочерних баз данных особенно важным является научное обоснование той информации, которая будет в ней накапливаться, чтобы избежать генерации никому не нужных баз. Поэтому созданию Asite предшествовал длительный период анализа. Можно дать следующую краткую характеристику информации по структуре активных сайтов.

Хорошо известно, что в основе реализации функциональной активности белков лежат специфические взаимодействия этих молекул с другими макромолекулами — мишенями. Области белков, участвующие в этих контактах называются активными сайтами. Информация о контактных областях, является фундаментом в решении любых задач, связанных с функциональной активностью белков. В PDB отведены специальные записи, куда заносится список атомов, образующих водородные связи при контакте. Однако эффективность контактов во многом определяется аминокислотами из локального окружения активных сайтов. Наличие структурных данных об областях в белках, включающих контактные аминокислоты и окружающие их аминокислоты, имеет чрезвычайно важное значение для молекулярной биологии. Следует заметить, что представление такой информации сопряжено с большими трудностями, так как контактные области, как правило, состоят из аминокислот, удаленных друг от друга по первичной последовательности. Представление таких аминокислот в виде ансамбля точек в пространстве может быть эффективно использовано в исследованиях, но предполагает дополнительную обработку данных в решении конкретных задач. Структурное выравнивание, идентификация структурных паттернов, выявление геометрической комплиментарности являются задачами высокой степени сложности и требуют интенсивных компьютерных расчетов. Поэтому, наряду с представлением активных сайтов как группы сближенных в третичной структуре, но разрозненных по первич-

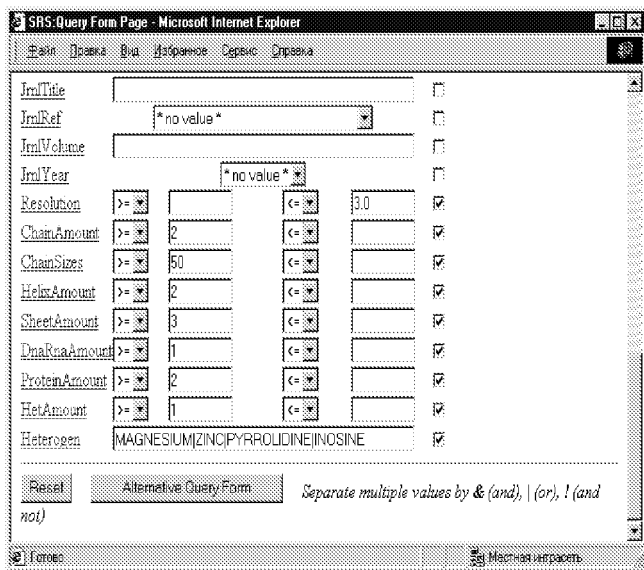


Рис. 6: Приведен пример запроса с заданными значениями количества цепей, длины цепей белка, а также количества альфа-спиралей, бета-структур, молекул нуклеиновых кислот, и список лигандов, которые должны присутствовать в искомым записях.

ной структуре аминокислот, мы разработали и иное представление. Оно основывается на представлении участка поверхности белка в виде набора пептидов заданной длины, состоящих из таких последовательностей аминокислотных остатков, которые удовлетворяют стереохимическим требованиям для полипептидных цепей, хотя могут и не быть связаны друг с другом ковалентными связями. В настоящее время имеются доказательства того, что такие пептиды могут функционально имитировать активные сайты [10]. Подобное представление дает возможность анализа поверхностных участков белков традиционными средствами анализа первичных последовательностей и имеет огромное количество потенциальных пользователей.

Алгоритм расчета такой информации включает в себя расчет доступной для растворителя поверхности белка, построение графа, описывающего связи между аминокислотными остатками данного участка поверхности белка и преобразование графа связности в набор линейных пептидных последовательностей.

На текущий момент времени нами разработаны необходимые программы анализа, формат представления данных в базе и ведется ее наполнение данными. Пример одной записи из базы Asite приведен на рис. 7. Расчет таких данных является длительным процессом, в связи с чем эта база пока не доступна по сети Интернет. Мы предполагаем к моменту публикации настоящей работы создать с использованием средств SRS Интернет-доступный вариант базы Asite, интегрированный с другими базами, в том числе и ресурсами GeneExpress.

## 5 База данных ASPD

В связи с очень быстрым ростом количества информации по аффинной селекции пептидов, на сегодняшний день

ID	A1	- идентификатор
SN	TOXIN	- научное название молекулы
SR	SEA SNAKE VENOM	- источник
DR	PDB:EWTA001S; SWISS: S29675;	- ссылки на комплементарные базы
PL	6	- длина пептидов
SN	12;29;35;64;71;72;73;124;225;	- список позиций акт сайта
SA	A;V;S;W;E;R;A;D;L;	- список аминокислот акт сайта
PN	12;35;71;72;73;124;	- позиции aa в пептидах
PN	12;71;72;73;124;35;	
PN	12;35;71;72;73;124;	
PN	71;72;73;124;225;29;	
PN	225;12;35;71;72;73;	
...		
PS	A;V;S;W;E;R;	- аминокислотная посл. пептидов
...		
...		

Рис. 7: Пример одной записи из базы Asite.

остро стоит задача ее накопления в виде базы данных, систематизации и обеспечения к ней Интернет-доступа, а также интеграции этой базы с другими биологическими информационными ресурсами.

База по аффинной селекции пептидов создана нами для накопления экспериментальных данных аффинной селекции пептидов, опубликованных разными авторами. В базе собирается информация об аминокислотных последовательностях селективированных пептидов, об особенностях молекулы-селектора, об аффинности и условиях экспериментов, при которых эти данные были получены. Такие данные представляют большой интерес для фундаментальных исследований взаимодействий между белком и различными лигандами, а также могут быть использованы в прикладных исследованиях по конструированию разного рода лекарственных, профилактических и диагностических пептидных препаратов.

Следует обратить внимание на то, что с развитием такой новейшей технологии аффинной селекции как фаговый дисплей в печати появились многие сотни и даже тысячи статей, содержащих такую информацию. Метод фагового дисплея широко применяется для различных целей, например, для выявления лигандов, активность которых *in vitro* может в несколько раз превышать активность природных белков, для определения разрывных эпитопов, опознаваемых антителами, для замены полипептидами небелковых лигандов [17]. Этот метод применялся и *in vivo* для определения белков, связывающихся в определенных тканях [12].

База данных по аффинной селекции пептидов в настоящее время доступна через систему SRS (рис. 8).

## 6 Пакет программ CRASP

Для изучения и анализа ко-адаптивных замен в аминокислотных последовательностях белковых семейств нами разработан пакет программ CRASP. Этот пакет программ позволяет выявлять и анализировать замены аминокислотных остатков в белках, которые происходят коррелированным образом. В основе анализа лежит предположение о том, что замены пар аминокислотных остатков белка, между которыми существуют функционально важные взаимодействия, в ходе эволюции происходят

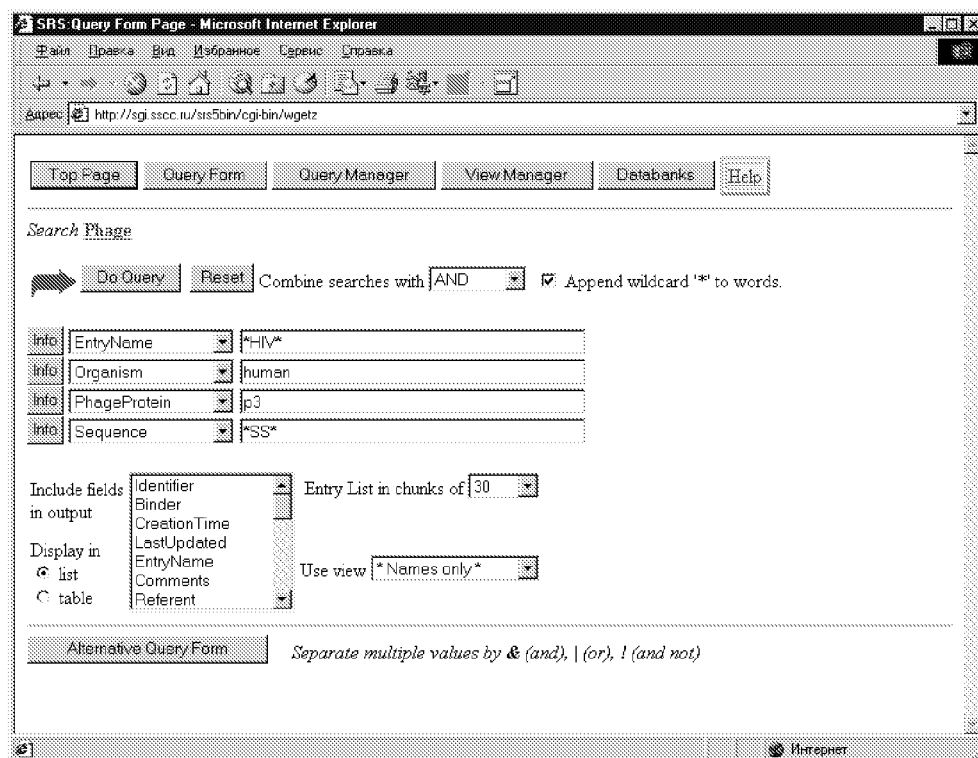


Рис. 8: Пример страницы запроса базы данных по аффинной селекции пептидов в формате системы SRS.

зависимым образом [3, 4]. Соответственно, обнаруженные статистические зависимости между заменами аминокислот в позициях белка могут указывать на существование функционально важных взаимодействий между остатками в этих позициях. В частности предполагается, что некоторые из обнаруженных корреляций могут иметь компенсаторную природу, обусловленную пространственной близостью остатков в коррелирующих позициях [3, 4, 8, 16]. В итоге, информация о координированных заменах может быть использована для предсказания контактов аминокислотных остатков в пространственной структуре белков. Такая информация может быть использована для выявления функционально важных взаимодействий в белковой глобуле, для планирования мутационных экспериментов в функциональных участках белка, а так же и для предсказания их пространственной структуры белка [11, 14].

Другой важной особенностью пакета программ CRASP является возможность выявления и изучения консервативных физико-химических характеристик белка, т. е. таких параметров, которые имеют тенденцию поддерживаться на некотором постоянном уровне в ходе эволюции белкового семейства. Исследованию подобного явления был посвящен ранее ряд работ, в которых оценивалось постоянство объема гидрофобного ядра белковой глобулы ряда белковых семейств [3, 7]. Результаты анализа позволили предположить, что подобные характеристики могут отражать важнейшие особенности структурно-функциональной организации белковых молекул.

Разработанный нами пакет программ позволяет:

- выявлять координированные аминокислотные за-

мены на основе оценки коэффициента корреляции между величинами определенного физико-химического свойства в паре позиций множественного выравнивания белков;

- анализировать полученную картину корреляций;
- выявлять как консервативные, так и переменные физико-химические характеристики белка;
- оценивать степень консервативности (вариативности) выявленных характеристик в сравнении с гипотезой независимых замен остатков в позициях белка.

Блок-схема пакета CRASP представлена на рис. 9.

Программный пакет CRASP имеет www-интерфейс и доступен пользователям через Internet по адресу [20]. Пакет представляет собой две группы программ: анализа парных корреляций аминокислотных замен и анализа интегральных характеристик групп позиций белка. Отметим, что как блок анализа парных корреляций, так и блок анализа функциональных характеристик могут быть использованы независимо друг от друга.

## 6.1 Основные методы и алгоритмы. Анализ парных корреляций замен аминокислот

Для выявления корреляций аминокислотных замен в позициях белка используется анализ физико-химических характеристик аминокислот, таких как объем боковой группы, заряд, гидрофобность и т. д. Эти анализируемые параметры выбираются пользователем из существующих

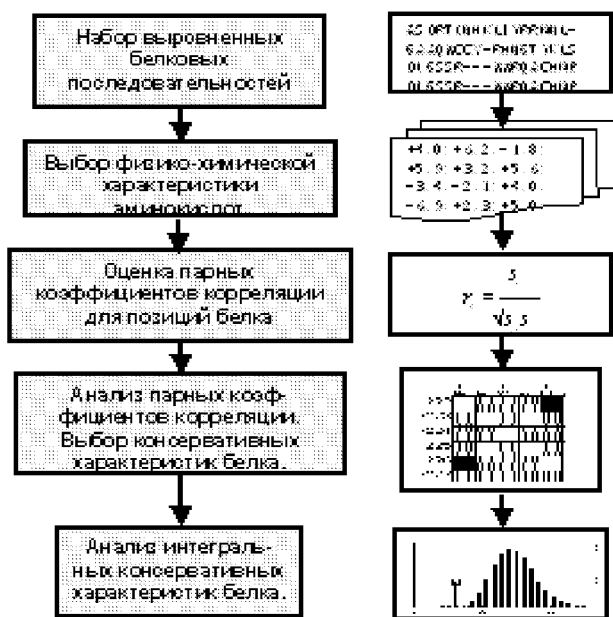


Рис. 9: Блок-схема пакета программ CRASP.

ющей базы данных физико-химических свойств аминокислот, которая в настоящее время содержит набор из 36 характеристик. Предполагается, что эти характеристики отражают взаимодействия остатков в белковой глобуле, поэтому выявленные зависимости между величинами физико-химических свойств аминокислот могут указывать на специфические функционально важные взаимодействия между остатками.

Для анализа используется набор выровненных аминокислотных последовательностей белкового семейства. При анализе каждый тип аминокислоты в матрице выравнивания заменяется на соответствующее значение выбранной физико-химической характеристики аминокислот. В качестве меры зависимости аминокислотных замен в позициях белка (столбцах полученной числовой матрицы) используется величина коэффициента корреляции между значениями физико-химической характеристики [1]. Чем ближе по абсолютному значению его величина к 1, тем сильнее статистическая зависимость аминокислотных замен для данной пары позиций белка. В пакете программ CRASP реализована так же возможность оценки частных коэффициентов корреляции, которые позволяют выявлять степень непосредственной зависимости двух позиций при условии, что аминокислоты в остальных позициях белка не изменяются [1]. Пакет программ позволяет так же использовать различные способы взвешивания анализируемых данных.

## 6.2 Анализ консервативных интегральных характеристик белка

В качестве интегральных характеристик белка  $F$  рассматриваются линейные комбинации физико-химических свойств аминокислот в позициях его последовательности. Эти характеристики отражают взаимодействия групп аминокислотных остатков как целого (активных сайтов

с лигандом, элементов вторичной структуры и других структурных элементов глобулы между собой). Консервативность такой интегральной характеристики может означать, что взаимодействия остатков в соответствующей группе позиций являются важными с точки зрения структуры и функции белка. В качестве меры консервативности величины  $F$  для выборки последовательностей белкового семейства используется ее выборочная дисперсия. Для выявления консервативных интегральных характеристик мы предлагаем использовать информацию о парных корреляциях величин физико-химического свойства в позициях выравнивания. В настоящее время этот этап анализа не автоматизирован и выбор интегральной характеристики оставляется за пользователем.

Для оценки степени консервативности характеристики  $F$  величина ее дисперсии сравнивается с о значением, ожидаемым в предположении независимых замен аминокислот в позициях белка. В пакете CRASP для оценки консервативности величины  $F$  используется так же процедура Монте-Карло. При этой процедуре с помощью случайных перестановок аминокислот в позициях выравнивания генерируется большое число выборок, таких, что распределение аминокислот в позициях выравнивания идентично реальным белкам, при этом возможные парные корреляции аминокислотных замен устраняются. В каждой такой выборке оценивается дисперсия функциональной характеристики  $D_{\text{rnd}}(F)$ . Далее значение  $D(F)$  реальной выборки сравнивается с распределением величин  $D_{\text{rnd}}(F)$  в случайных выборках. В частности оценивается вероятность того, что значение  $D(F)$  будет меньше ожидаемого для выборки последовательностей с независимыми заменами аминокислот. Эта вероятность вычисляется как отношение числа случайных выборок, в которых  $D_{\text{rnd}}(F) > D(F)$  к полному числу сгенерированных выборок. Отметим, что расчеты методом Монте-Карло требуют больших затрат машинного времени, поэтому в Internet версии пакета CRASP число случайных выборок ограничено величиной 10000. При анализе последовательностей средней длины (около 60 позиций белка) и размерах выборки около 100 время расчета составляет около 2–3 минут. Результаты анализа показывают, что оценки, выполненные двумя этими методами оказываются весьма близкими, различие составляет несколько процентов и меньше.

Пакет программ CRASP предполагает одновременный анализ до 4 интегральных характеристик белка и позволяет оценивать их дисперсии, взаимные корреляции, а так же параметры линейной регрессии. Результаты могут быть представлены как в текстовом виде, так и в виде графических диаграмм.

## 6.3 Реализация пользовательского интерфейса

Пакет программ CRASP реализован на языке C. Пользовательский интерфейс реализован в виде HTML-форм (рис. 10). Этот интерфейс позволяет пользователю вводить набор анализируемых данных, устанавливать параметры численных расчетов и определять режим вывода полученных результатов.

## 6.4 Входные данные

Основными входными данными для анализа являются набор выровненных родственных последовательностей белкового семейства. Кроме того, предоставляется возможность изменения следующих параметров расчета:



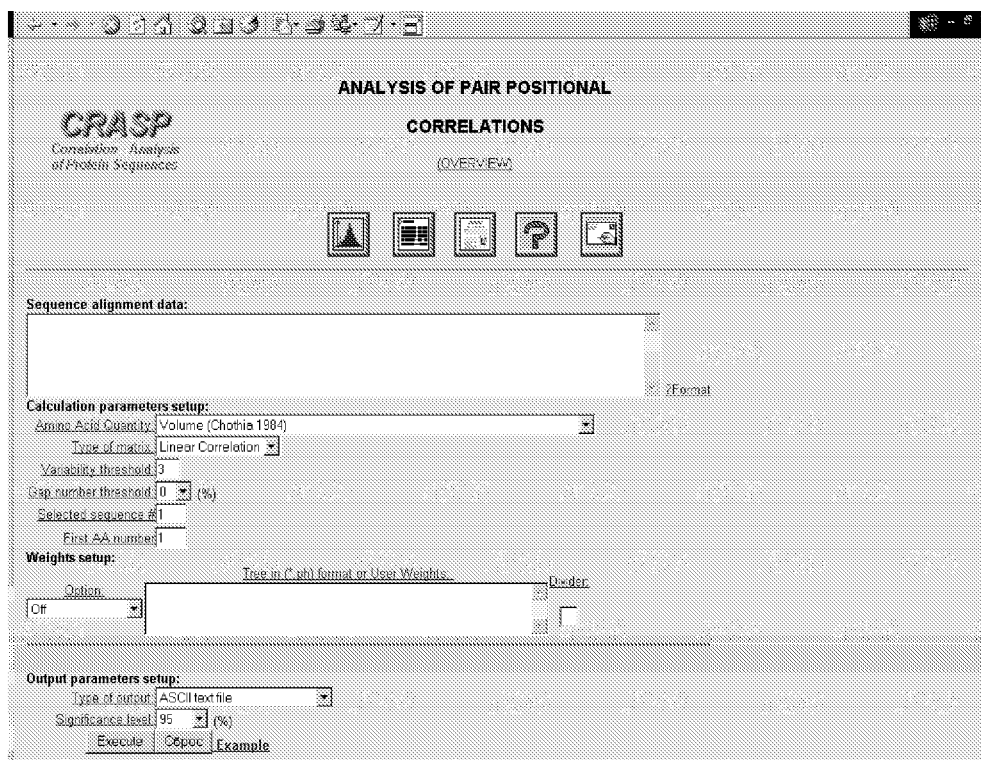


Рис. 10: Пример WWW-страницы для программы анализа парных корреляций.

- Выбор физико-химической характеристики для анализа;
- Установка порога варируемости позиций белка, включаемых в анализ.
- Установка порога на число делеций для позиций, включаемых в анализ. Делеции рассматриваются как “пропущенные данные”. При оценке величин дисперсий значение такой “пропущенной” физико-химической характеристики заменяется на среднюю по столбцу выравнивания. Подобная замена, естественно, приводит к искажениям в оценке параметров распределения, поэтому верхний порог на число делеций в столбце выравнивания ограничен величиной в 15 % от числа анализируемых последовательностей.
- Установка нумерации позиций исследуемых последовательностей;
- Выбор схемы взвешивания данных и ввод параметров взвешивания.

При анализе парных корреляций необходимо так же указать тип матрицы зависимости позиций, которая будет рассчитана.

При анализе интегральных характеристик белка пользователь может задавать до 4 характеристик, указывая номер позиции белка и численный коэффициент, с которым значение физико-химической характеристики в этой позиции входит в интегральную характеристику белка.

## 6.5 Параметры вывода результатов анализа

Полученные в результате расчетов данные представляются как набор из нескольких взаимосвязанных HTML-документов. Это позволяет в удобном для пользователя виде представлять результаты анализа белковых последовательностей.

Для программы анализа парных корреляций это:

- Вывод матрицы статистической зависимости пар позиций белка. Этот вывод по желанию пользователя может осуществляться в виде текстового документа, в виде HTML-таблицы, с выделением цветом наиболее значимых коэффициентов корреляции или в виде графической диаграммы (рисунок в формате GIF).
- Вывод диаграммы парной зависимости позиций белка в виде бинарного дерева.
- Вывод в текстовом формате параметров распределения значений физико-химической характеристики в позициях последовательности (варируемость позиции, среднее значение и дисперсия характеристики).
- Вывод распределения коэффициентов корреляции (в текстовом формате).

Для программы анализа интегральных характеристик белка это:

- Вывод значений интегральных характеристик для белков из анализируемой выборки. Существует возможность вывода этих данных в текстовом виде, а

так же в виде графических диаграмм (GIF-файлы): гистограммы распределений отдельных характеристик в выборке и диаграммы рассеяния пар значений исследуемых характеристик в белках.

- Вывод результатов сравнения дисперсий исследуемых характеристик в выборке реальных белков со значениями, ожидаемыми для случайных выборок. Этот вывод возможен как в текстовом виде, так и в виде гистограмм.

Отметим, что система CRASP имеет гипертекстовый файл помощи с описанием используемых алгоритмов и параметров расчетов. Поэтому в любой момент пользователь имеет возможность получить необходимую подсказку. На сайте приведены так же образцы заполнения HTML-форм и наборов анализируемых данных.

## 6.6 Пример анализа белков с помощью системы CRASP

С помощью системы CRASP нами был проведен анализ ряда белковых доменов, которые взаимодействуют с ДНК.

При анализе доменов класса гомеодомов оказалось, что замены в нескольких позициях белка значимо коррелируют по величине изоэлектрической точки аминокислот (характеристика, отражающая заряд аминокислоты). На диаграмме, представляющей парные корреляции позиций в виде бинарного дерева, это группы позиций 19, 30, 33 и 15, 18, 37, обозначенные как кластеры **A** и **B** (рис. 11). Аминокислотные остатки в позициях 15, 18, 19 расположены на поверхности двух альфа-спиралей домена (H1 и H2). При анализе пространственной структуры белка оказалось, что остатки в этих позициях оказываются сближенными таким образом, что между ними возникают электростатические взаимодействия (солевые мостики), стабилизирующие укладку двух альфа-спиралей гомеодомена (рис. 12). Это взаимодействия разноименно заряженных остатков белка. Отрицательные значения коэффициентов корреляции величины заряда в этих позициях, обнаруженные нами, свидетельствуют, что замены аминокислот в этих позициях белка происходят так, что взаимодействия разноименных зарядов сохраняются, а полный заряд остатков в этих позициях поддерживается на постоянном уровне.

В целом же, система взаимодействий для этих двух кластеров позиций, по-видимому, представляет собой набор из нескольких комплиментарных солевых мостиков, которые не являются локализованными в какой-то одной паре позиций, а могут вовлекать несколько соседних позиций и формируются динамически, в зависимости от характера замещений остатков в каждой конкретной последовательности белка в ходе его эволюции.

Полный заряд остатков в районе контактов двух альфа-спиралей H1 и H2 поддерживается в ходе эволюции на некотором оптимальном уровне. Его дисперсия почти в полтора раза меньше, чем это ожидается для случайных последовательностей гомеодоменов (15.02 и 23.41, соответственно).

Эти результаты подтверждаются анализом диаграммы рассеяния величин  $Q_{H1}$  (суммарный заряд в позициях 15, 18, 19 спирали H1) и  $Q_{H2}$  (суммарный заряд в позициях 30, 33, 37 спирали H2), приведенной на рис. 13. Из рисунка видно, что распределение значений этих величин в анализируемой выборке белков следует линейной

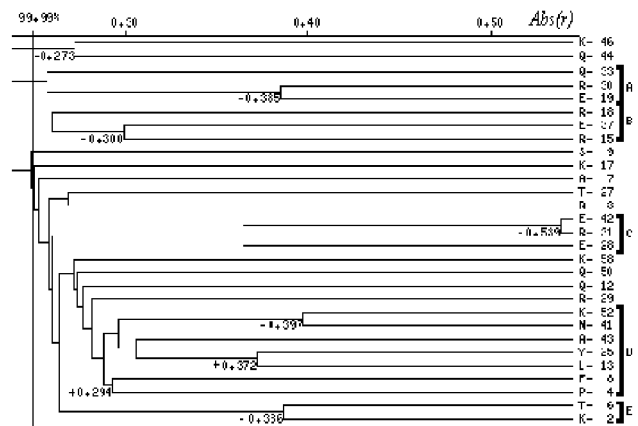


Рис. 11: Представление позиций гомеодомена, замены в которых коррелируют по величине изоэлектрической точки аминокислот, в виде бинарного дерева. Справа приведены обозначения позиций гомеодомена. Вверху приведена шкала абсолютного значения коэффициента корреляции. Критические значения коэффициента корреляции (при различных уровнях значимости) показаны вертикальными линиями. Для некоторых пар позиций приведены численные значения коэффициента корреляции. Кластеры из наиболее коррелирующих позиций обозначены латинскими буквами справа.

зависимости с отрицательным наклоном линии регрессии. Это означает, что чем большее значение принимает величина  $Q_{H1}$ , тем меньшее значение в данном белке имеет величина  $Q_{H2}$ , что является отражением компенсаторного по величине заряда эффекта замен.

При анализе другого класса ДНК-связывающих белков, содержащих структурный мотив типа “лейциновая заставка” мы обнаружили эффект сохранения полного заряда аминокислотных остатков на концах ДНК-связывающих альфа-спиралей этих белков [1]. Консервативность этой характеристики по сравнению со случайными наборами последовательностей хорошо заметна из рис. 14. Дисперсия этой величины существенно меньше, чем соответствующие значения для случайных белков. Полученные результаты хорошо согласуются с имеющимися экспериментальными данными и позволили нам предложить схему взаимодействий между зарядами на концах распознающих спиралей и ДНК [1].

В целом, использование подобного подхода позволяет получать важную информацию о взаимодействиях в структуре белка на основании анализа набора родственных последовательностей белкового семейства.

## 7 Заключение

Ранее GeneExpress был ориентирован на описание функций только ДНК и РНК. Разработанная нами система FrameProt стала важной частью системы GeneExpress, обеспечивая пользователей базовой информацией по структуре и функциям трех типов биологических макромолекул — ДНК, РНК и белков.

Планы дальнейшего ее развития включают наращивание методов анализа, создание новых баз данных по

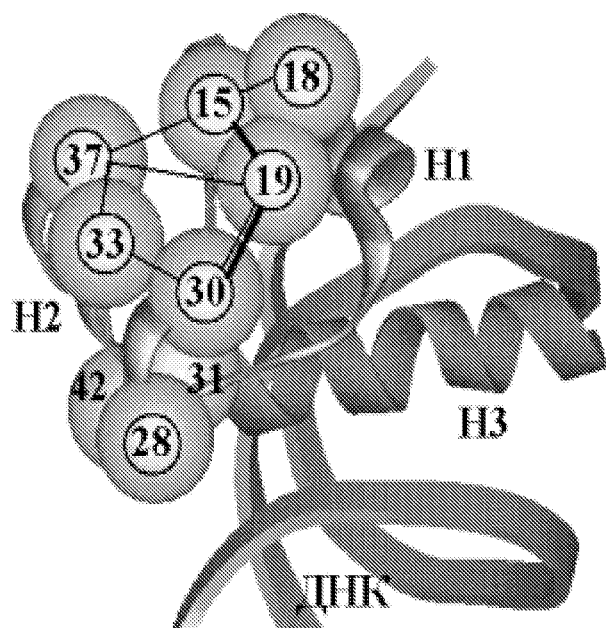


Рис. 12: Расположение остатков в позициях кластеров А, В, С в пространственной структуре гомеодомена. показаны пары коррелирующих остатков (тонкие линии) и солевые мостики (жирные линии). Альфа-спирали гомеодомена обозначены как H1, H2, H3. Спираль H3 ложится в большую бороздку ДНК, образуя с ней контакты, две других спирали параллельны между собой и лежат поперек спирали H3.

структурным и функциональным особенностям белков, создание баз знаний и усовершенствование интеграции с другими информационными биологическими ресурсами.

Следует подчеркнуть, что PDB является ежедневно обновляемой базой данных. Поэтому с целью поддержания целостности создаваемых ресурсов в рамках проекта планируется динамическое обновление базы EnPDB, а также дочерних баз данных. Разработанная нами программа конвертации из PDB в EnPDB делает эту задачу выполнимой в ближайшем будущем.

Возможности поиска EnPDB будут расширены путем добавления новых полей, содержащих информацию о принадлежности белка к тому или иному типу пространственной укладки так называемому фолдингу.

Наличие большого количества баз данных, интегрированных в единую информационную систему, очень остро ставит задачу создания эффективных систем запросов для поиска содержащейся в ней информации и различных манипуляций с ней (классификация, сортировка и т. д.). Решение этой задачи требует нетривиальных подходов, так как не может быть осуществлено средствами только стандартных языков запросов. Одним из наиболее эффективных подходов является создание проблемно-ориентированного языка высокого уровня, обеспечивающего решение этой задачи на основе семантического анализа информации, содержащейся в совокупности баз данных. В настоящее время в лаборатории Теоретической Генетики ИЦиГ СО РАН ведется разработка такого языка — MGL [2]. Планируется расширить функ-

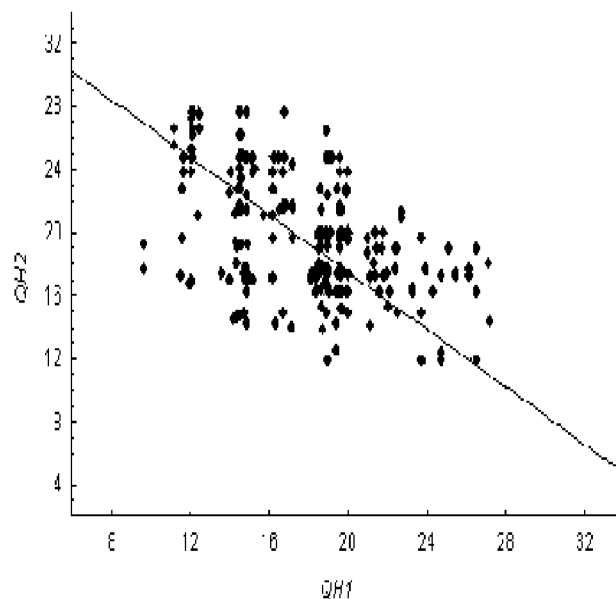


Рис. 13: Зависимость величин  $Q_{H1}$  от  $Q_{H2}$  для исследованных белков. Определения величин  $Q_{H1}$  и  $Q_{H2}$  даны в тексте. Коэффициенты корреляции и уравнения регрессии для этой зависимости:  $r = -0.51$ ,  $Q_{H2} = 35.66 - 0.91 \cdot Q_{H1}$ .

циональные возможности этого языка для применения в области пространственных белковых структур.

Будут разработаны такие программы анализа пространственных структур белков, как структурное выравнивание белковых поверхностей, активных сайтов и будет проведена их интеграция с EnPDB.

Предполагается дальнейшее развитие пакета CRASP. Важным этапом развития этой системы является ее интеграция с базой данных пространственных структур белковых макромолекул EnPDB. Эта интеграция обеспечит возможность получения информации об особенностях пространственного расположения коррелирующих позиций, таких как расположение пар остатков относительно активных сайтов белка, доступность их растворителю, принадлежность различным типам вторичной структуры белка и т. д. Структурная информация о белке будет отображаться совместно с информацией о парных корреляциях аминокислотных замен в позициях белка. Это позволит пользователю эффективно оценивать функциональную нагрузку коррелирующих позиций.

Предполагается так же произвести интеграцию пакета CRASP с существующими базами данных множественных выравниваний белковых последовательностей. Это позволит утилизировать огромное количество информации о последовательностях белков, накопленной в настоящее время и использовать ее совместно с информацией о структурах этих белков. Результаты анализа множественных выравниваний белков различных семейств предполагается представить в виде новой информационной базы данных.

Выполнение поставленных задач позволит создать информационный ресурс, необходимый широкому кругу

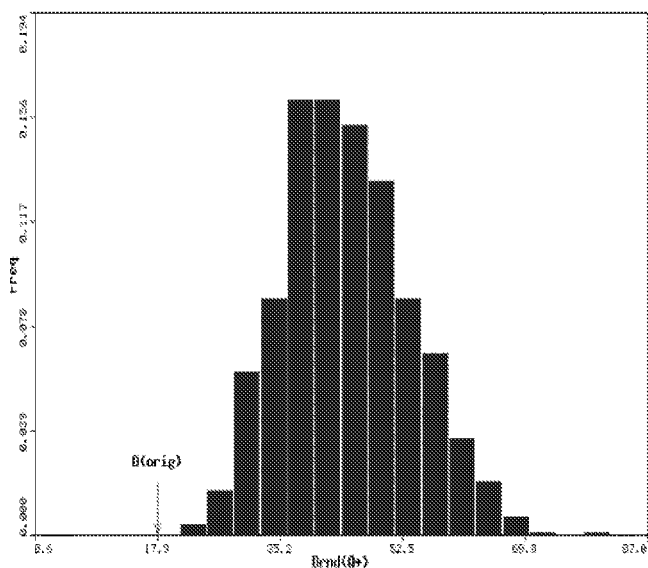


Рис. 14: Пример графического вывода программы анализа интегральных характеристик белка. Значение дисперсии суммарного заряда ДНК-связывающих спиралей белков, содержащих структурный мотив типа “лейциновая застежка” (показано красной стрелкой) и распределение соответствующих величин для 1000 случайных последовательностей (гистограмма).

специалистов в области молекулярной биологии и белковой инженерии.

Работа поддержана грантами РФФИ 97-04-49740, 98-04-49479, 98-07-90126, 98-07-91078, 99-07-90203.

## Библиография

- [1] Афонников Д.А., Кондрахин Ю.В., Титов И.И. *Выявление коррелирующих позиций ДНК-связывающего района транскрипционных факторов семейства CREB и AP-1.* // Мол. Биол. 1997. Т. 31. С. 741-748.
- [2] Колчанов Н.А. *ГЕНЭКСПРЕСС: электронная библиотека по структурам и функциям ДНК, РНК и белков.* // Доклады конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”. Санкт-Петербург. 1999.
- [3] Лим В.И., Птицин О.Б. *О постоянстве объема гидрофобного ядра в молекулах миоглобинов и гемоглобинов.* // Мол. Биол. 1970. Т. 4. С. 372-382.
- [4] Altshuh D., Lesk A.M., Bloomer A.C., Klug A. *Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus.* // J. Mol. Biol. 1987. V. 193. P. 693-707.
- [5] Barker W.C., Garavelli J.S., Haft D.H., Hunt L.T., Marzec C.R., Orcutt B.C., Srinivasarao G.Y., Yeh L.S.L., Ledley R.S., Mewes H.W., Pfeifer F., Tsugita A. *The PIR-international protein sequence database.* // Nucleic Acids Res. 1998. V. 26. P. 27-32
- [6] Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D., Rodgers J.R., Kennard O.,

Shimanouchi T., Tasumi M. *The Protein Data Bank: a computer based archival file for macromolecular structures.* // J.Mol.Biol. 1977. V. 112. P. 535-542.

- [7] Gerstein M., Sonnhammer E.L.L., Chothia C. *Volume changes in protein evolution.* // J. Mol. Biol. 1994. V. 236. P. 1076-1078.
- [8] Gobel U., Sander C., Schneider R., Valencia A. *Correlated mutations and residue contacts in proteins.* // Prot. Struct. Funct. Genet. 1994. V. 18. P. 309-317.
- [9] Kolchanov N.A. *Transcription Regulatory Regions Database (TRRD): its status in 1999.* // Nucleic Acids Res. 1999. V. 27. P. 303-306.
- [10] Luzaggo A., Felici F., Tramantano A., Pessi a., Cortese R. *Mimicking of discontinuous epitopes by phage-displayed peptides. 1. Epitope mapping of human ferritin using a phage library of constrained peptides.* // Gene. 1993. V. 128. P. 51-57.
- [11] Ortiz A.R., Kolinski A., Skolnick J. *Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignment.* // J. Mol. Biol. 1998. V. 277. P. 419-448.
- [12] Pasqualini R. and Ruoslahti E. *Organ targeting in vivo using phage display peptide libraries.* // Nature. 1996. V. 380. P. 364-366.
- [13] Patikoglou G., Burley S.K. *Eukaryotic transcription factor-dna complexes.* // Annu. Rev. Biophys. Biomol. Struct. 1997. V. 26 P. 289-325.
- [14] Pazos F., Helmer-Citterich M., Ausiello G. and Valencia A. *Correlated mutations contain information about protein-protein interaction.* // J. Mol. Biol. 1997. V. 271. P. 511-523.
- [15] Ponomarenko M.P. *Generating programs for predicting the activity of functional sites.* // J. Comput. Biol. 1997. V. 4. P. 83-97.
- [16] Shindyalov I.N., Kolchanov N.A., Sander C. *Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations?* // Prot. Eng. 1994. V. 7. P. 349-358.
- [17] Smith G.P. *Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface.* // Science. 1985. V. 228 P. 1315-1317.
- [18] [<http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/>]
- [19] [<http://srs5.bionet.nsc.ru/srs5/>]
- [20] [<http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp/>]