

# Использование SQL для совместного поиска на химических данных, белковых последовательностях и трехмерных структурах

Адель Головин



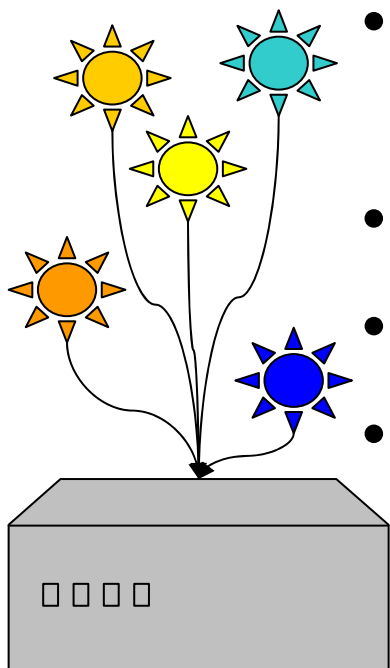
# PDB – источник информации

- PDB это богатство и единство представления данных о ~63.000 структур
  - Детали экспериментальная
  - Белковые и ДНК последовательности ~48.000
  - Маленькие молекулы и элементы цепочек ~10.000
  - Параметры кристаллизации, кристаллы
  - 3D координаты атомов
- Вторичные структуры, 3D мотивы, PROSITE, взаимодействия молекул
- CATH, PFAM, PRINTS, ... все доступное через DAS (Distributed Annotation System)



# Distributed Annotation System

DAS <http://www.dasregistry.org/listServices.jsp>



- Единый формат запроса и ответа для всех серверов
- Центральный регистр: 659 серверов
- Две системы координат: UNIPROT, PDB
- UNIPROT:
  - PFAM, PIRSF, PRINTS, PRODOM, PROSITE, ...
- PDB:
  - CATH, SCOP, EC, Secondary structure



# Стимулы для разработки

- Помочь биологическим исследованиям
  - Находить сходства и расхождения
  - Находить склонности и зависимости
  - Открыть новое знание и обрести понимание
- Нужен всесторонний (1-2-3)M поиск
  - Независимый от платформы
  - Масштабируемый по процессорам, памяти, дискам
  - Справляющийся с большими объемами данных
  - Open source
  - На SQL



# SQL как основное средство

- Почему на SQL?
- Это возможно и естественно
  - RDBMS созданы решать сетевые задачи
  - Сама реляционная база это гипер-сеть
  - SQL запрос это подсеть
- Достоинства:
  - Легко расширять и встраивать
  - Широкий выбор поисковых движков (MySQL, PostgreSQL, DB2, Oracle, Sybase...)
  - Решения на базе SQL приходят пакетом с управлением данными и ресурсами



# SQL для химических подструктур

- Таблицы: compounds, atoms, bonds

- Пример 1: O=C-N

```
Select b1.a1, b1.a2, b2.a2
from bonds b1
join bonds b2 on b2.a1= b1.a2
where b1.type=X1 and b2.type=X2
```

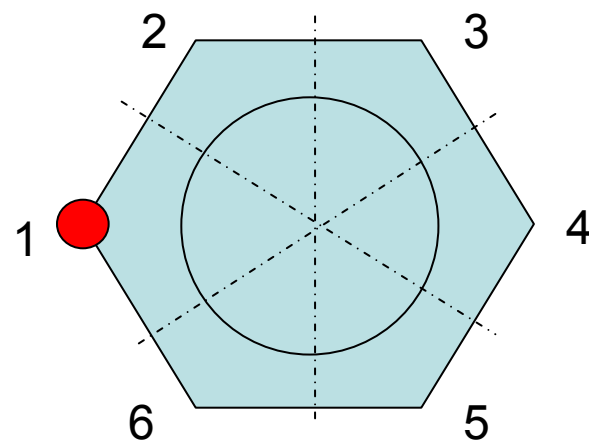
- Пример 2: C-C-C

```
Select b1.a1, b1.a2, b2.a2
from bonds b1
join bonds b2 on b2.a1= b1.a2
where b1.type=X and b2.type=X and b1.a1 != b2.a2
```



# Различать себя и отражение

- Путь в обход всех перестановок
- Бензиновое кольцо
- 12 перестановок атомов
- $a_1 < a_4$ ,  $a_1 < a_3$ ,  $a_1 < a_5$ ,  $a_2 < a_6$
- Включить это в SQL
- <http://www.ebi.ac.uk/pdbe-site/chemsearch/>
- Constraint satisfaction problem



# Pattern search in SQL

- $N\{P\}[ST]\{P\}$  : N-Glycosylation центр
- Последовательные значения в колонке ID таблицы базы данных для элементов последовательностей ( $id2 = id1 + 1$ )
- SQL: 

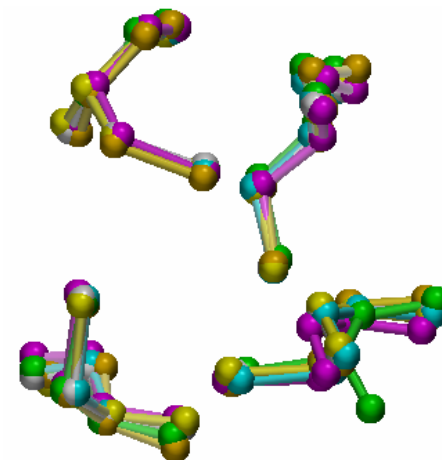
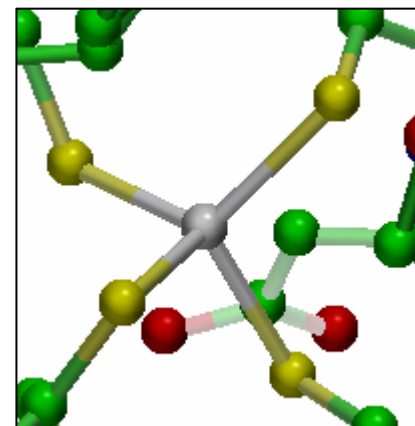
```
select r1.id, r4.id from residues r1
      join residues r2 on (r2.id = r1.id + 1)
      join residues r3 on (r3.id = r2.id + 1)
      join residues r4 on (r4.id = r3.id + 1)
where r1.code='N' and r2.code!='P' and
      r3.code in ('S','T') and r4.code!='P'
```
- Тот же прием используется для последовательностей  $\phi/\psi$  углов в белках



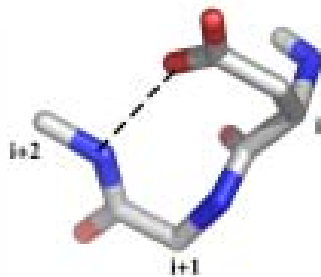
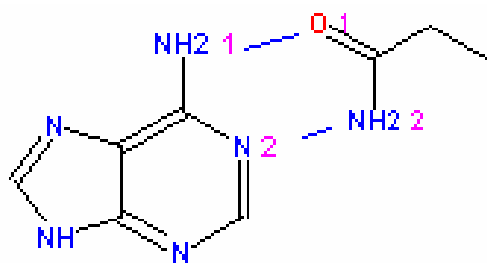


# Поиск 3М полостей на SQL

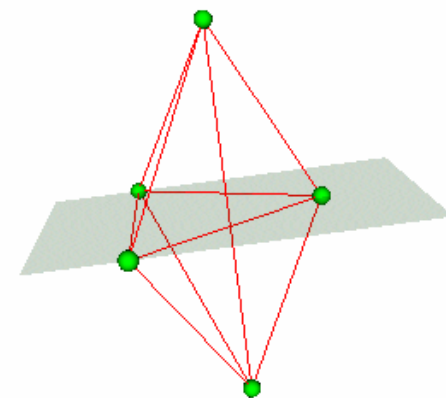
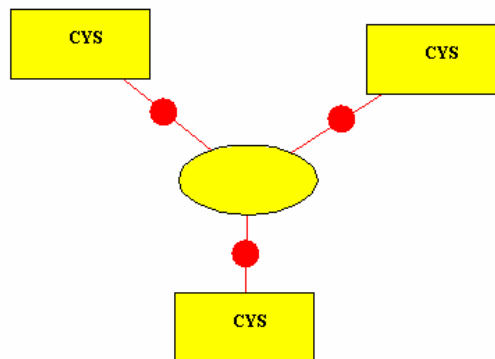
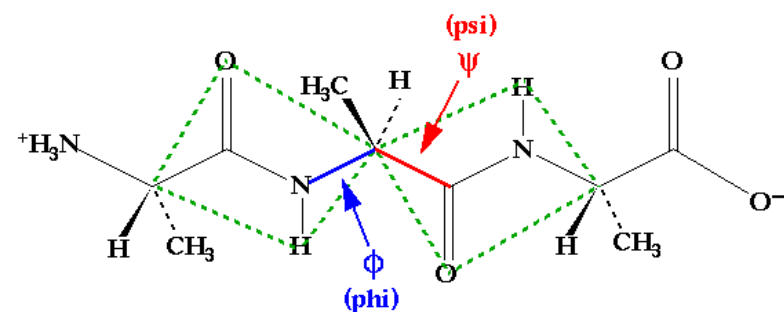
- Тот же подход как для химического поиска
- $C\alpha$  -  $C\alpha$  расстояния
- + Между активными атомами
- 16 Å ограничение
- Большинство запросов представляют полную сеть



# Широкий выбор элементов поиска



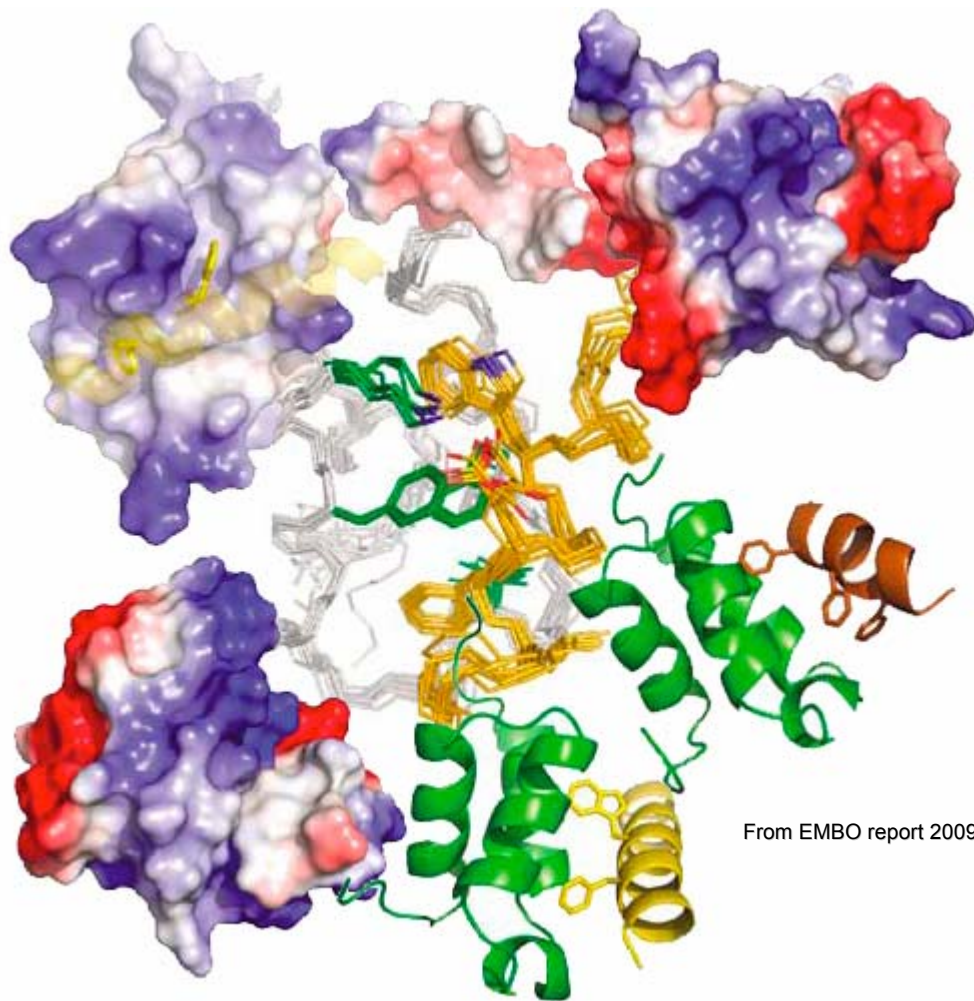
Human	V	S	K	P	L	G	P	A	S	F
Baboon	V	S	K	P	L	V	P	A	S	F
Bushbaby	A	V	K	P	L	V	P	A	S	L
Sheep	A	S	K	P	L	V	P	A	S	V
Cow	V	S	K	P	L	V	P	A	S	F
Pig	V	S	K	P	L	V	P	A	S	F
Marsupial	G	D	S	P	K	M	P	V	S	N



Совмещение всех элементов в одном SQL запросе



# Сеть взаимодействий



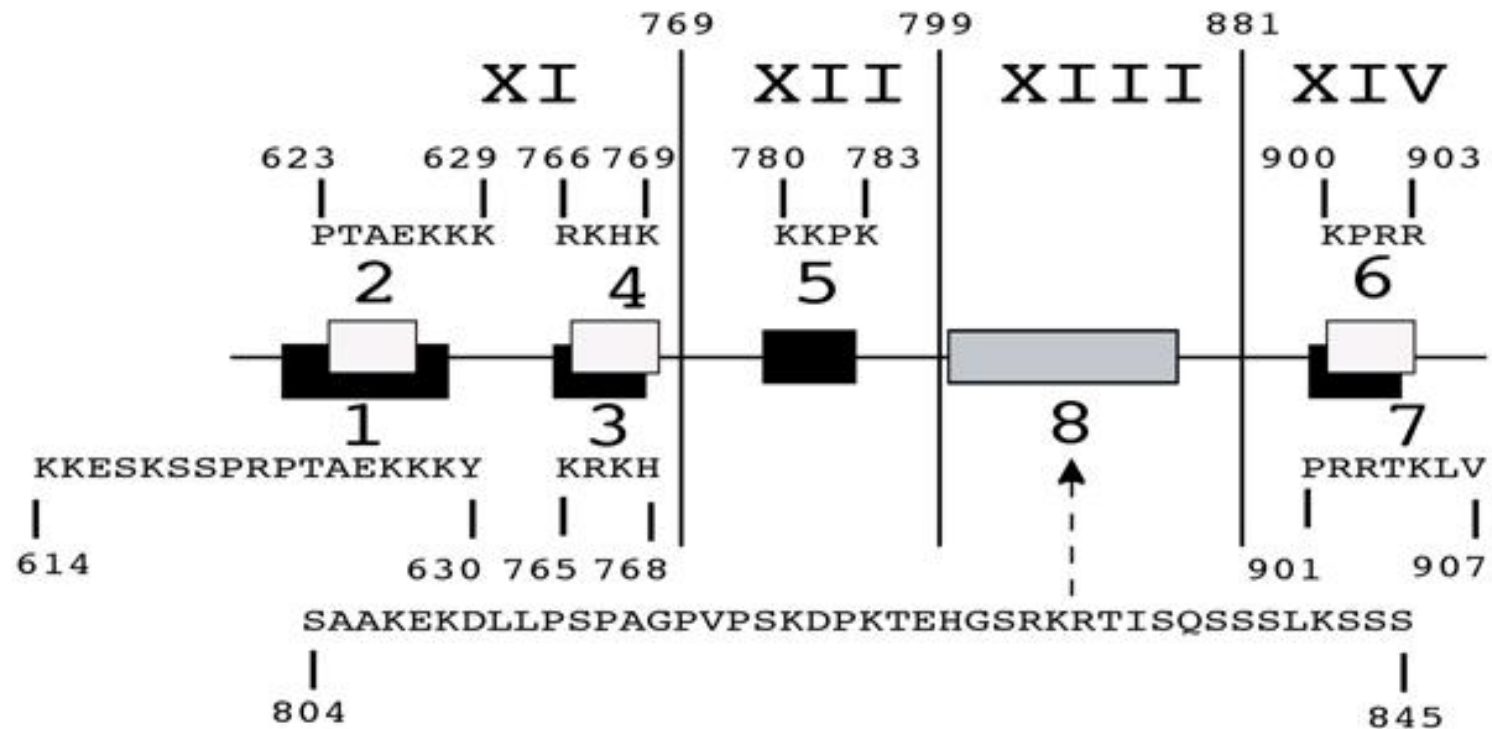
From EMBO report 2009

Взаимодействия:

- Ковалентные, ионные, электростатические
- Водородные
- Пи электронные



# Положение в белках/РНК/ДНК



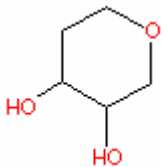
# Пример запроса

Вторичная структура - заколка

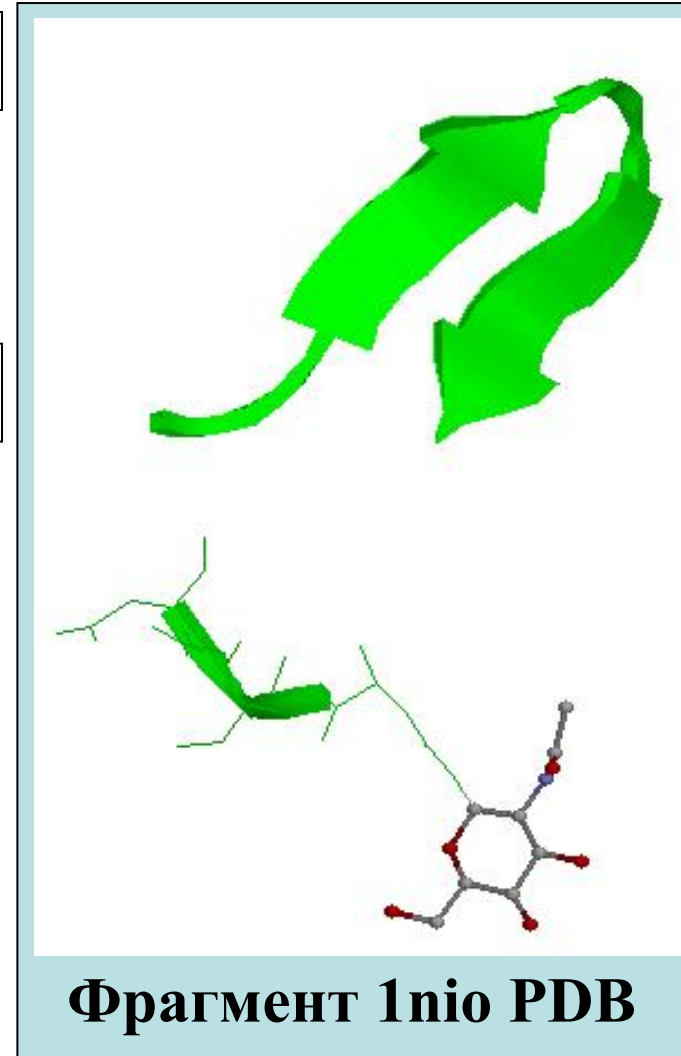
+ 2-3 аминокислоты пробел

+ N Glycosylation pattern N{P}[ST]{P}

+ Азот вступает в ковалентную связь в сахаридам:



Просматриваем ~170.000 моделей

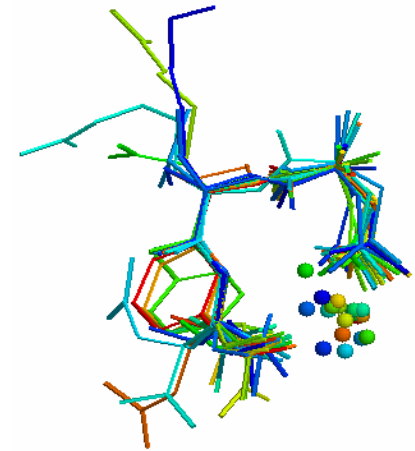


# Представление найденного

aligned by 50% sequences identity  
 search: sequence: GSGVKVAVLDTGISHTPDLNIRGGASFVPGFSTQDGNHGHTHVAGTIAALNNSIGVLGVAFSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMHVANLSLCL  
 filter: \*

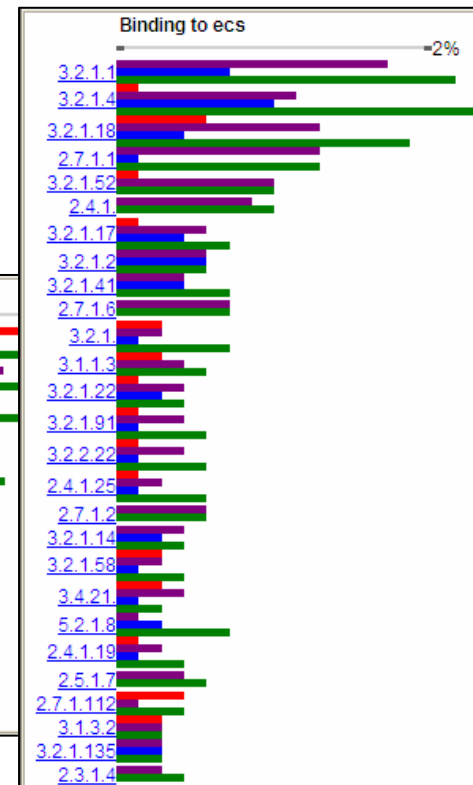
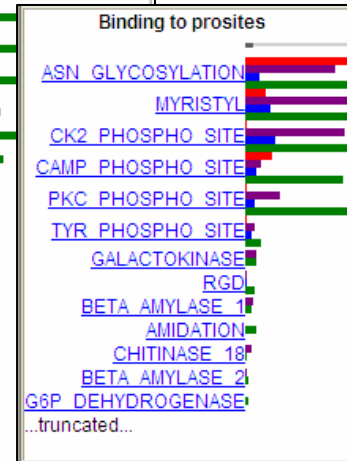
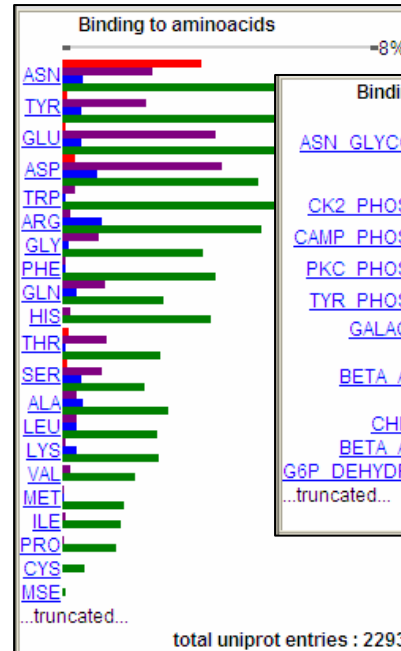
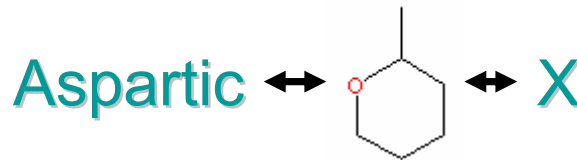
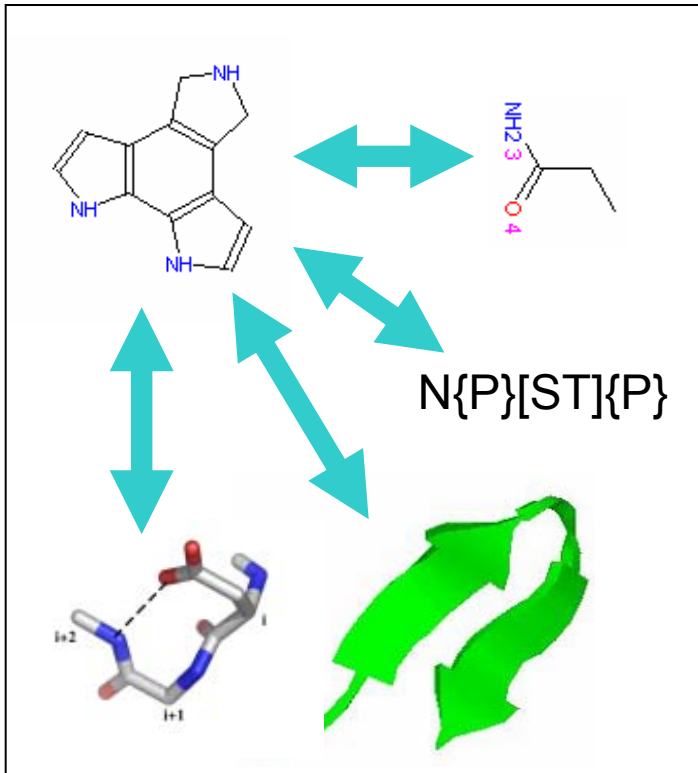
sequence	score
3SGVKVAVLDTGISHTPDLNIRGGASFVPGFSTQDGNHGHTHVAGTIAALNNSIGVLGVAFSAELYAVKVLGASGSGSVSSIAQGLEWAGNNGMHVANLSLCL	440.0
3SGAKIATVDTGVS-N-----HP--DL--NIRG-----GASF-----V---PGE--P-ST-Q-----DGNHGHTHVAGTIAAL--NNS--IQVGLVAPSAELY-----AVKVLGASGSGSVSSIAQGLEWAGNNGMHVANLSLCL	204.0
3SGIINAVLDTGV-N-T-----SHP--DLRNVVQ-----GDF-----TGATTPIN--N-SC-T-----DRNHGHTHVAGTIALD--GSSDQAGIYGVAPALV-----AYKVLDSGSGSVSSIAQGLEWAGNNGMHVANLSLCL	143.0
3FOVTAVYDITGV-N-N-----NHE--EF--GGRS-----VSGYDF-----V---DND--A-DS-S-----DGNHGHTHVAGTITGO-----SOGVAKNVNIV-----QVRVLSGSGSGITSGVIGVDVVAONASGSP-----VANMSLG	133.0
3TNIIVAVVDITGV-DOT-----HP--DL--EGQV-----IAGYRPAFDEEL--PAG--T-DS-S-----YGGSAQTHVAGTIAA--KGG--KGIYGVAPGAKIMPVIVFDPPALVGGNGYVDDVVAAGIIVATDH--GAK-----VMNHSWG	131.0
AGNKTICILDSGY-DRS-----HN--DL--NANNVTGNNSTGNVY-----Q---PG-----NNAHGHTHVAGTIAA--ANN--EGVGVGMPNQNAN-----IHIVKVFNEAGVYSSSLVAIDTCV--NSG-----GANVVM	106.0
3QGSQVVYDITGIEA-S-----HP--EF--E--G-----RAQH-----V---KTY--YYS-R-----DGNHGHTHVAGTIVGS-----RTYVAKTKTQF-----QVKVLDNNSGQYSTIAGMDFVASD--KNRNCPKGVVAVLSLCL	104.0
3HIVVSLDIDT-ERM-----HP--FLAGNYDF-----GASF-----D---VNDGP-DF--GPKYTOHNRKSTRGAGEVAVA--AMG--VGVGVAYAKRIG-----QVRMLDGLVDAVEARSLGLNPNHIIH--TYS-----ASWQPD	97.0
3GGIVAVADTGL-D--TORNDSSMHE--AF--RQKI-----TALY-----A---LGR--T--NRAN-----DGNHGHTHVAGSV--L-GNG--STNKQMAPGANLV-----FQSLMDSGGLOGLP--SLNGLFSGAY--SAG-----ARITHNS	92.0
3ISSYLNWYQKPK-G-K-----AP--KL--LHA-----ASSL-----E---TGV--P-SR-F-----SOSGOTDPSFTISL--OPE--DLATYQQQYDSL-----PLTFGGGKVEIKRIVAAPSVFIFPP--SDE-----OLKSGTA	58.0
... SESSESKA-S-S-----GL--PI--DLRG-----KRAF-----I---AG-----I--A-----DDNGYGVAVAKSLAA--GAE--ILVGTVPALNIF-----ETSLRRGKFDQSRVLPDGLSMEIKKV--YPL-----DAVDNP	54.0
3HOTLKNLTSVLT-S-V-----AD--TY--GKKV-----MVAE-----T---SYT--Y-TA-E-----DGGHGTAPKNGQTL--NNE--VIVQGGANAVRDV-----IQAVSDVGEAGIOVYVEPAWIPVGE--AHR-----LEKNKAL	52.0
KEAAEASREYLKI-A-C-----HP--ET--GLAP-----EYAY-----Y---DGT--P-ND-E-----KQYGHFFSDSVRYAAN--IGL--DAEVFGGSEWSAE-----EINKIQAFADKEPEDYRYKIDGEP--FEE-----KSLHPVG	48.0
... OFKDYGH-D-Y-----HPAKTE--NIKG-----LGLD-----K---PGI--P-KT-P-----KNGGGRKRVTDGK--RKI--YEVDSQHELEGY-----RASDGHLSGDFPKTONGLKOPDKR--NIK-----KYL	48.0

- Выравнивание последовательностей
- 3M выравнивание
- 3M варианты молекул
- Детальное представление белковых комплексов
- Частоты связывания и включения



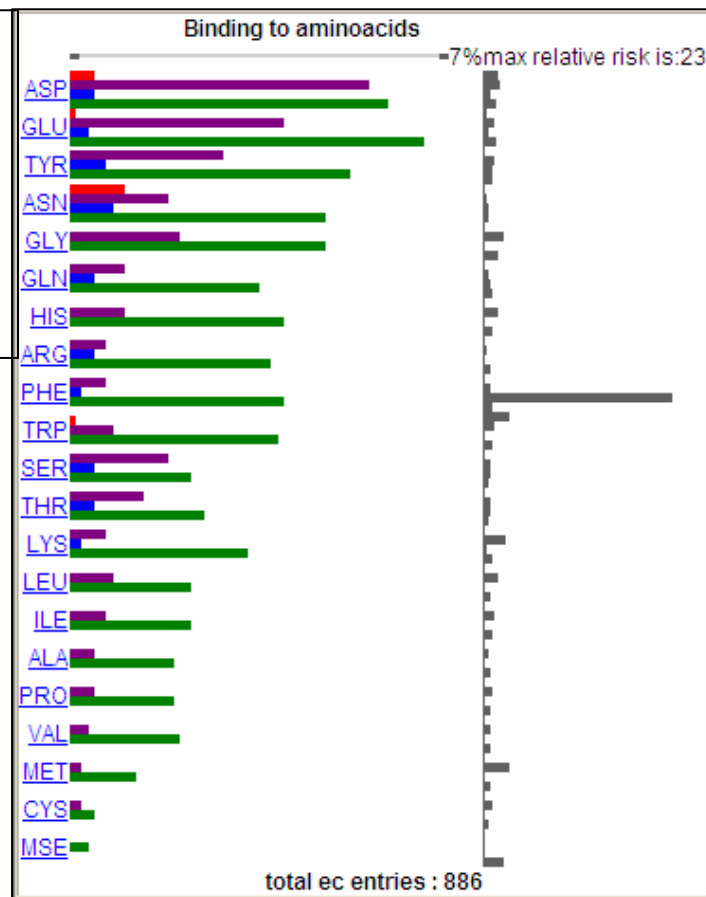
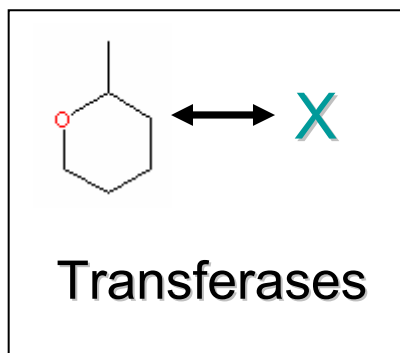
# Частоты связывания

Взаимодействие с аминокислотами, семьями белков, доменами, мотивами, активными центрами



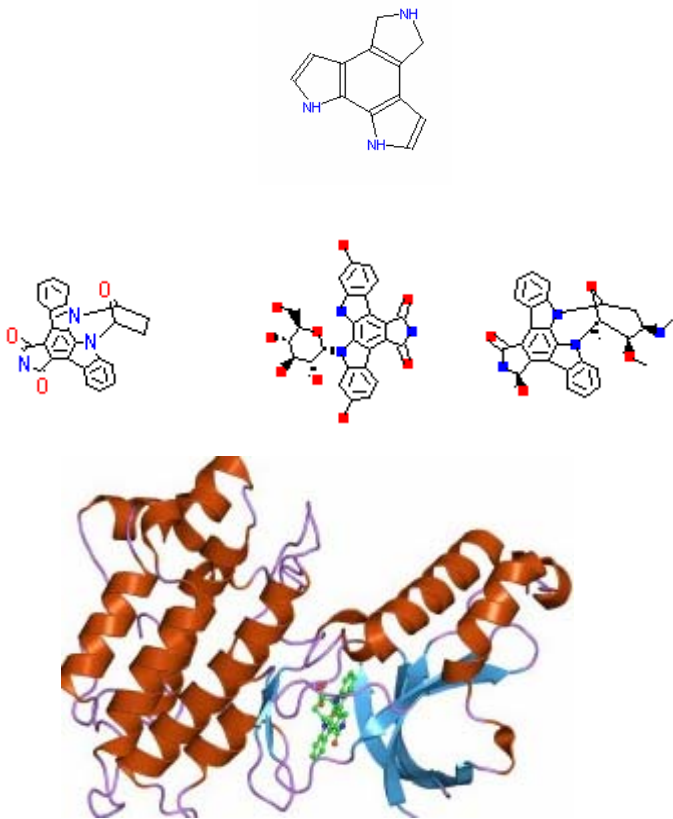
# Относительный риск

- Частоты на данной семье белков
- EC, CATH, PFAM, ...
- Какие взаимодействия выделяют эту семью?





# Свойства не PDB молекул в PDB



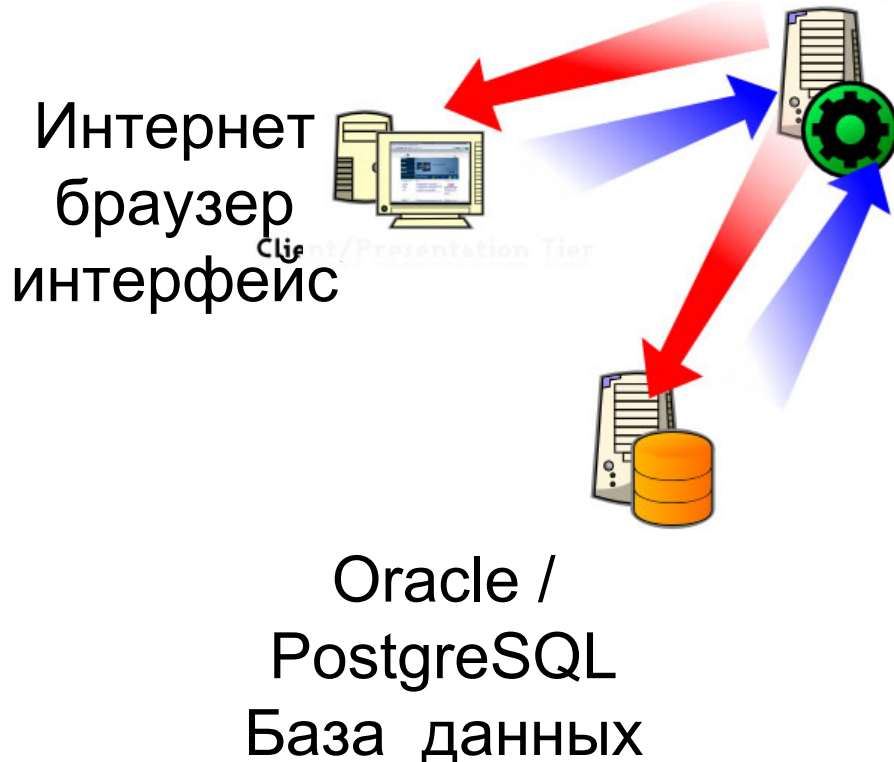
- Подструктура PDB молекул
- 3M варианты молекул
- Взаимодействия только подструктуры
- Соответствующие мотивы, активные центры и домены белков
- Дальнейшая связь с генетикой

```
VHFNEVIGRGHFGCVYHGLLD----KIHCAVKSLNRITDIGEVSQFLTEGIIMKD  
IELGRCIGEGQFGDVHQGIYMSPENPALAVAIKTKNCTSDSVREKFLQEALTMRQ  
LQLIKRLGNGQFGEVWMGTWNGNTK----VAIKTLKPGTMS--PESFLEEAQIMKK
```



# Наброски системы

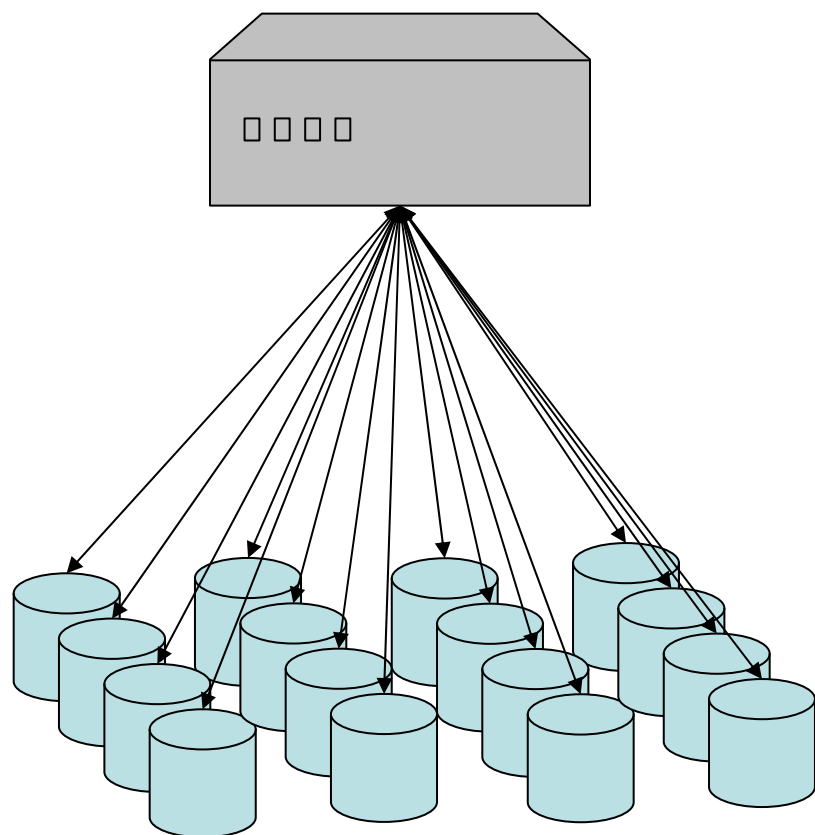
J2EE сервер приложений  
(weblogic, websphere, tomcat)



- Классическая n-уровневая
- Поддержка основных браузеров
- Поддержка разных движков баз данных



# N схем базы данных



16 схем / баз данных

60.000 PDB разделен на 16 частей

- Данные распределены между 16 схемами
- Преимущество по сравнению с партициями
  - Таблицы фактов (словари) тоже уменьшены в 16 раз
  - Нет зависимости от движка
  - Гарантия параллельных вычислений
- Сервер приложений
  - Одновременно 16 запросов
  - Объединяет и агрегирует
- Линейное ускорение
- Масштабируемость



# База данных

- «Разделяй и властвуй»
- Основные данные
  - Подразделены на таблицы белков, РНК/ДНК, молекулы окружения (лиганды), растворители (вода)
  - К PDB добавлены атомы водорода
- Словари:
  - Уникальные химические компоненты
  - Белковые и РНК/ДНК последовательности
  - PROSITE мотивы
  - Маленькие 3М мотивы ( $\beta$  – turns, ...)
  - Активные и каталитические центры, все из DAS



# Роль XML

- XML запрос – промежуточный уровень
- XML частотный запрос
- Выброс данных в XML
  - DAS, eFamily, родной XML
- PDB страницы существуют только в XML
- Легко встроить: AJAX, Google Kit, Yahoo Kit



# URLs

- “Chemical substructure search in SQL” Chem. Inf and Modelling, Jan 2009
- <http://www.ebi.ac.uk/pdbe-site/pdbemotif/>
- Частоты связывания для SMILES:  
[http://www.ebi.ac.uk/pdbe-site/pdbemotif/smilesstats.jsp?smiles=Nc1ncnc2\[nH\]cnc12](http://www.ebi.ac.uk/pdbe-site/pdbemotif/smilesstats.jsp?smiles=Nc1ncnc2[nH]cnc12)
- XML запрос:  
<http://www.ebi.ac.uk/pdbe-site/pdbemotif/hitlist.xml>  
<http://www.ebi.ac.uk/pdbe-site/pdbemotif/start?tab=help&topic=xml>
- Проект доступен (GPL):  
PDBSAM: <http://sourceforge.net/projects/pdbsam>  
CHEMSEARCH: <http://sourceforge.net/projects/chemsearch>



# Дело наше особое

- PDBeMotif – востребован:
  - ~1000 разных пользователей в неделю
  - ~30 стран активных пользователей
  - Количество активных пользователей растет
- EMBL безоговорочно согласно передать интеллектуальную собственность
- Адель Головин и Мелфорд Джон в конце 2010 года покидают EMBL чтобы открыть путь разработанной технологии

# Наша сила = наша слабость

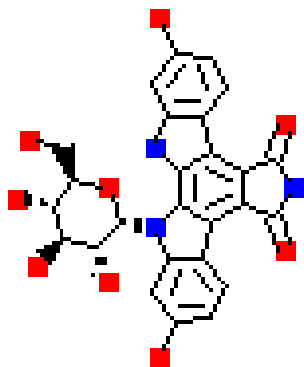
- 9 лет в академической среде = отрыв от реальности
- 9 лет внимания на одном продукте = потеря экспертизы в других областях
- 9 лет в Англии = отчуждение от России
- ИМБХ = моя последняя нить
- Этот доклад = **анти** рекламная акция



# Нужен совет

- Чем программа полезна ВАМ?
- В каких практических задачах может быть задействована?
- Какие есть сценарии использования?
- Что нужно доработать?
- Что нужно добавить?

# Пример: подроживание докингу



- Загрузить молекулу
- Пометить точки связывания
- Использовать 3-мерный поиск чтобы найти потенциально связываемые полости в белках из PDB
- Используется в автоматическом режиме для больших библиотек молекул
- Результаты отправляются в другую программу

# Нужен совет

- Чем программа полезна?
- В каких практических задачах может быть задействована?
- Какие есть сценарии использования?
- Что нужно доработать?
- Что нужно добавить?

[adel@frontlinesql.co.uk](mailto:adel@frontlinesql.co.uk)