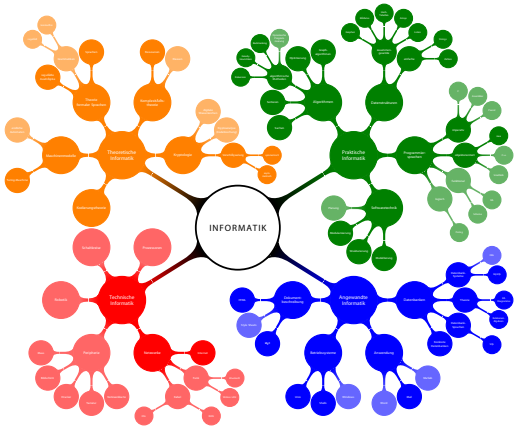
UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
THEORETISCHE INFORMATIK

Kapitel 42

Biodatenbanken

Wohin mit den Genomen?

Vorlesung Einführung in die Informatik 2 vom 1. Juli 2014 von Till Tantau



Lernziele von Kapitel 42

1. Überblick über das Feld der Biodatenbanken bekommen
2. Arten von Biodatenbanken kennen
3. Grenzen und Möglichkeiten von Biodatenbanken einschätzen können
4. Auf Daten in Beispieldatenbanken zugreifen können

Gliederung von Kapitel 42

Was sind Biodatenbanken?

Was sind Biodatenbanken?

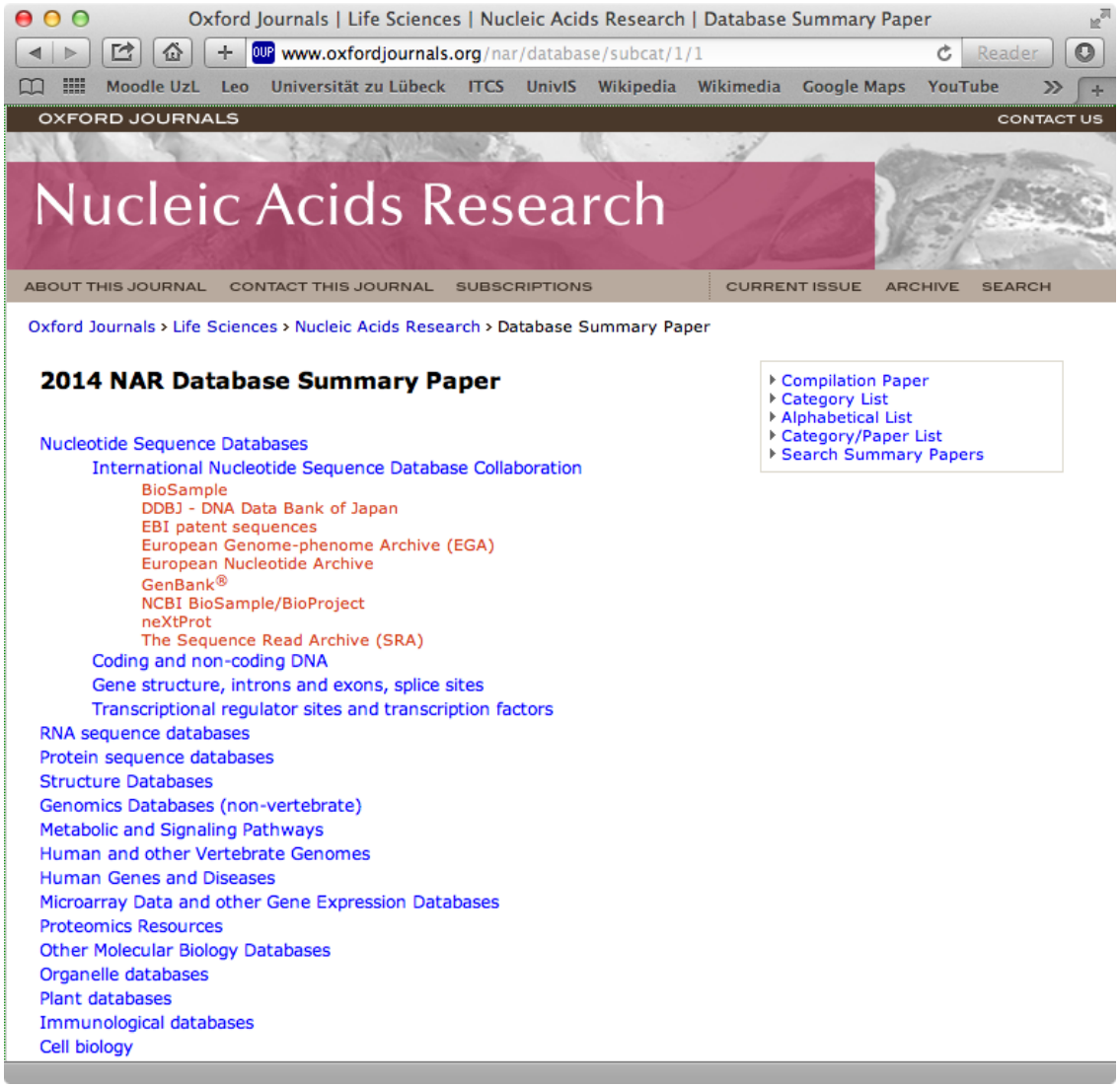
- Eine *Biodatenbank* ist eine digitalisierte *Sammlung von biologischen Daten*.
- Es muss sich *nicht* um eine relationale Datenbank handeln; selbst eine Ansammlung von Textdateien kann eine Biodatenbank darstellen.

Liste der Biodatenabanken

- 1000 Genomes Selection Browser
- 16S and 23S Ribosomal RNA Mutation Database
- 2D-PAGE
- 2P2Idb
- 3D rRNA modification maps
- 3D-Footprint
- . . .
- ZiFDB
- ZInC

Es gibt viele Biodatenbanken.

Die Anzahl der Biodatenbanken ist so groß (über 1500 im Juni 2014), dass es eine Datenbank dieser Datenbanken gibt auf http://www.oxfordjournals.org/nar/database/cap:



Klassifikation von Biodatenbanken.

Man kann Biodatenbanken klassifizieren nach der *Art der gespeicherten Daten*:

- DNA-Sequenz-Datenbanken
- RNA-Sequenz-Datenbanken
- Protein-Sequenz-Datenbanken
- Molekülstruktur-Datenbanken
- Gen-Datenbanken
- Genexpressions-Datenbanken
- Pathway-Datenbanken
- Publikations-Datenbanken

Von jedem Typ gibt es sehr viele Datenbanken ganz unterschiedlicher Art, Umfang und Zielsetzung.

Was speichert eine Biodatenbank?

Rohdaten versus aufbereitete Daten

Eine Biodatenbank kann zwei Arten von Dingen speichern:

- Die *Rohdaten*, die in Untersuchungen gewonnen wurden. Dies schließt Informationen über das *wer, wie, was, wo, warum* betreffend die Experimente ein.
  Beispiel: Bei einem Microarray-Experiment die Fotos des Arrays.
  Beispiel: Bei einer Sequenzierung die ermittelten Fragmente.
- *Aufbereitete Daten,* die aus den Rohdaten gewonnen wurden.
  Beispiel: Bei einem Microarray-Experiment verschiedene ermittelte Gencluster.
  Beispiel: Bei einer Sequenzierung der prognostizierte »Golden Path«.

Aus welchen Quellen speist sich eine Biodatenbank?

Rohquellen versus aufbreitete Quellen

Die Daten in einer Biodatenbank können auf zwei Arten gesammelt werden:

- Forschungsgruppen *»submitten«* Daten zu einem Thema bei einer Biodatenbank, die diese dann allen Interessenten zugänglich macht.
- Die Betreiber der Biodatenbank *sammeln* die Daten bei verschiedenen Forschungsgruppen regelmäßig nach eigenen Kriterien ein, bereiten diese auf und stellen dann das Ergebnis allen Interessenten zur Verfügung.

Zur Diskussion

Welche Vor- und Nachteile haben die Verfahren?

Arten, wie Daten in Biodatenbanken gespeichert werden.

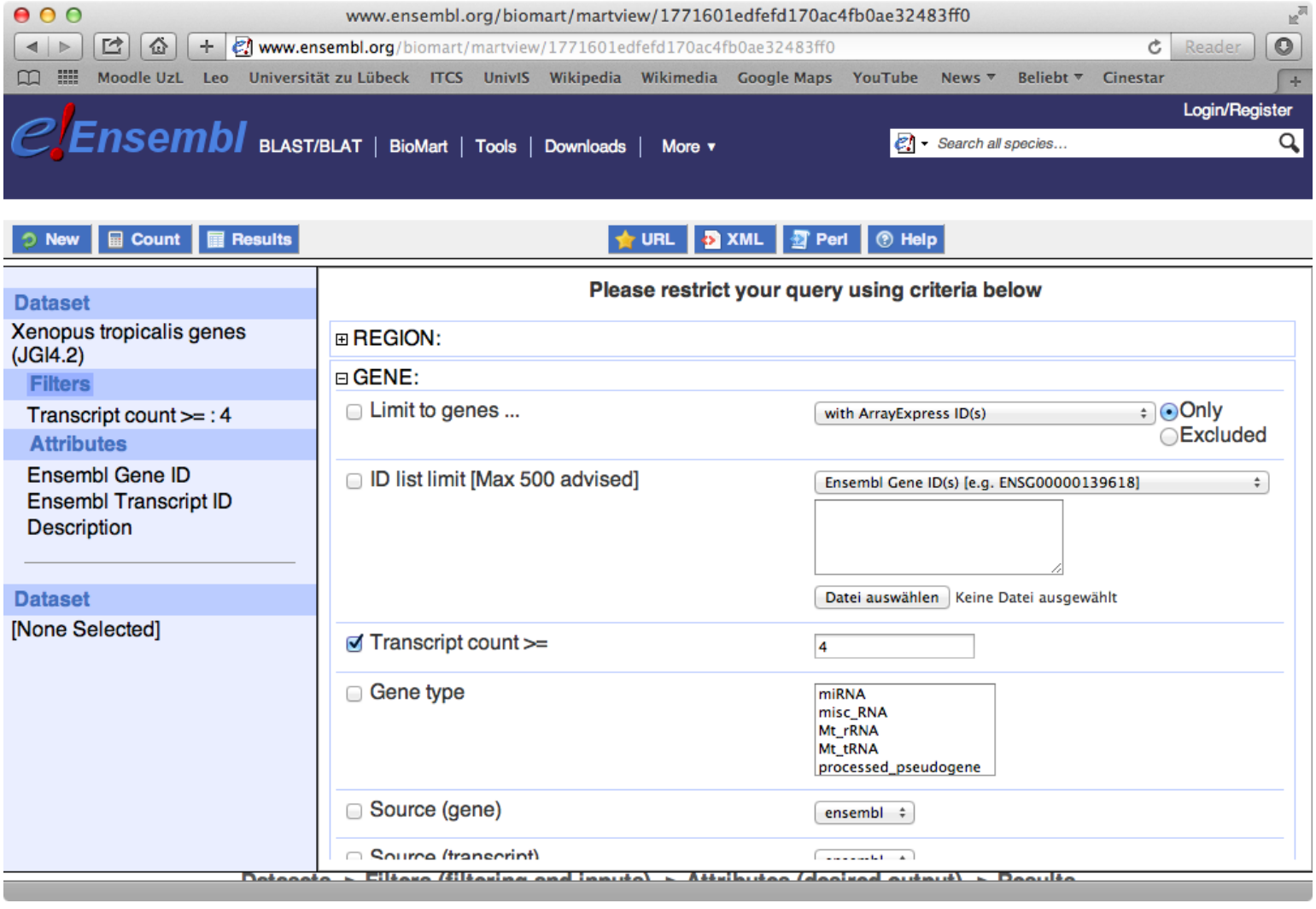
Die in einer Biodatenbank gespeicherten Daten können unterschiedlich stark strukturiert sein. Mögliche *Datenmodelle* sind:

1. *Reiner Text:* Beispiele sind Sammlungen von Artikeln zu einem Thema.
2. *Entry-basiert:* Lange Listen von Zeilen (Entries), die eine spezielle Syntax haben (beispielsweise mit einem Schlüsselwort wie `ATOM` oder `REMARK` anfangen).
3. *Relationale:* Die Datenbank ist eine SQL-Datenbank mit einem Entity-Relationship-Schema.
4. *Objektorientiert:* Die Datenbank ist eine OO-Datenbank.
5. *XML:* Die Daten werden als XML-Dateien entsprechend einem bestimmten Schema gespeichert.

Wer kann wie auf die Daten zugreifen?

- Bei *öffentlichen* Biodatenbanken kann erstmal jeder auf die Daten zugreifen.
- Es gibt in der Regel ein *Web-Interface,* über das man Daten finden kann und
- oft auch komplexere Anfragen stellen kann.
- Unter Umständen gibt es auch einen direkten Zugang zu einem SQL-Server.

Wer kann wie auf die Daten zugreifen?



Interne Speicherung versus externer Zugriff

- Jede Biodatenbank speichert Daten intern auf eine bestimmte Art.
- Bei einem *externen Zugriff* kann man auf die Daten aber meist auf mehrere Weisen herunterladen; die Daten werden »zur Not umgerechnet«.
- Zur Verfügung stehen:
  - Flat-Files,
  - XML-Dateien und machmal
  - SQL-Zugriff auf die Datenbank.

Biodaten als Flat-Files

- Ein *Flat-File* ist eine einfache Textdatei, in der jede Zeile einen kleinen Datensatz darstellt.
- Es gibt keinerlei Standard für die Formate, jede Biodatenbank nutzt typischerweise selber mehrere »selbst ausgedachte« Formate.

```
HEADER    HYDROLASE/HYDROLASE INHIBITOR           01-SEP-11   3TNS
TITLE     CRYSTAL STRUCTURE OF SARS CORONAVIRUS MAIN PROTEASE COMPLEXED WITH AN
TITLE    2 ALPHA, BETA-UNSATURATED ETHYL ESTER INHIBITOR SG83
...
ATOM      1  N   SER A   1      23.652   4.764 -24.058  1.00 24.36           N
ANISOU    1  N   SER A   1     2645   3238   3373   -593    126   -557       N
ATOM      2  CA  SER A   1      22.698   5.744 -23.500  1.00 24.91           C
ANISOU    2  CA  SER A   1     2800   3292   3372   -616    100   -508       C
```

Biodaten als XML-Dateien

- Bei *XML* handelt es sich um einen Standard, *wie man Daten strukturiert und aufschreibt*. (Siehe auch das Kapitel hierzu.)
- Man kann mit einer *Document Type Definition* genau und recht »sauber« festlegen, welche Tags in einer Datei vorkommen dürfen und was sie bedeuten.
- Generell lassen sich XML-Dateien von Computern leichter als Flat-Files bearbeiten, sind aber eher groß.

```
<?xml version="1.0" encoding="UTF-8" ?>
<PDBx:datablock datablockName="3TNS"...>
   <PDBx:atom_siteCategory>
      <PDBx:atom_site id="1">
         <PDBx:B_iso_or_equiv>24.36</PDBx:B_iso_or_equiv>
         <PDBx:Cartn_x>23.652</PDBx:Cartn_x>
         <PDBx:Cartn_y>4.764</PDBx:Cartn_y>
         <PDBx:Cartn_z>-24.058</PDBx:Cartn_z>
         <PDBx:auth_atom_id>N</PDBx:auth_atom_id>
         <PDBx:auth_comp_id>SER</PDBx:auth_comp_id>
         ...
      </PDBx:atom_site>
      <PDBx:atom_site id="2">
         <PDBx:B_iso_or_equiv>24.91</PDBx:B_iso_or_equiv>
         <PDBx:Cartn_x>22.698</PDBx:Cartn_x>
         <PDBx:Cartn_y>5.744</PDBx:Cartn_y>
         <PDBx:Cartn_z>-23.500</PDBx:Cartn_z>
         <PDBx:auth_atom_id>CA</PDBx:auth_atom_id>
         ...
```

Zugriff auf SQL-Datenbanken

- Bei den meisten Biodatenbank steht heutzutage »zumindest im Hintergrund« eine SQL-Datenbank.
- Manche Biodatenbanken geben jedem sofort (Ensembl) oder zumindest auf Anfrage einen direkten Lese-Zugriff auf die Datenbank.

```
> mysql --host=ensembldb.ensembl.org --user=anonymous
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 3872065
...
mysql> use xenopus_tropicalis_core_75_42;
```

Die Protein-Data-Bank: www.rcsb.org

- Die Protein-Data-Bank wird von den Universitäten Rutgers und UCSD betrieben.
- Sie speichert *Molekülstrukturdaten,* die beispielsweise aus NMR-Spektroskopien entstanden sind.
- Der Zugriff erfolgt in zwei Schritten:

1. Über das Webinterface und dessen Suchfunktionen findet man die Identifikationsnummer (ID) des interessierenden Proteins.
2. Mit dieser Nummer kann man dann die Struktur des Proteins als Flat-File oder als XML-Datei herunterladen.

Alternativ kann man auch gleich »im Browser« das Molekül betrachten.

Das Flat-File-Format der Protein-Data-Bank


```
HEADER    HYDROLASE/HYDROLASE INHIBITOR           01-SEP-11   3TNS
TITLE     CRYSTAL STRUCTURE OF SARS CORONAVIRUS MAIN PROTEASE COMPLEXED WITH AN
TITLE    2 ALPHA, BETA-UNSATURATED ETHYL ESTER INHIBITOR SG83
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SARS CORONAVIRUS MAIN PROTEASE;
...
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: SARS CORONAVIRUS;
SOURCE   3 ORGANISM_COMMON: SARS-COV;
...
KEYWDS    3C-LIKE PROTEASE, PROTEASE, HYDROLASE-HYDROLASE INHIBITOR COMPLEX
EXPDTA    X-RAY DIFFRACTION
AUTHOR    L.ZHU,R.HILGENFELD
REVDAT   1   05-SEP-12 3TNS    0
JRNL        AUTH   L.ZHU,R.HILGENFELD
JRNL        TITL   CRYSTAL STRUCTURES OF SARS-COV MAIN PROTEASE COMPLEXED WITH
JRNL        TITL 2 A SERIES OF PEPTIDIC UNSATURATED ESTERS
...
ATOM      1  N   SER A   1      23.652   4.764 -24.058  1.00 24.36           N
ANISOU    1  N   SER A   1     2645   3238   3373   -593    126   -557       N
ATOM      2  CA  SER A   1      22.698   5.744 -23.500  1.00 24.91           C
ANISOU    2  CA  SER A   1     2800   3292   3372   -616    100   -508       C
```

Wie das Format funktioniert


```
HEADER    HYDROLASE/HYDROLASE INHIBITOR           01-SEP-11   3TNS
TITLE     CRYSTAL STRUCTURE OF SARS CORONAVIRUS MAIN PROTEASE COMPLEXED WITH AN
TITLE    2 ALPHA, BETA-UNSATURATED ETHYL ESTER INHIBITOR SG83
...
ATOM      1  N   SER A   1      23.652   4.764 -24.058  1.00 24.36           N
```

Das Flat-File-Format der PDB folgt einer festen Struktur

- Jede Zeile hat maximal 80 Zeichen und enthält keine Sonderzeichen und keine Umlaute.
- Die ersten sechs Zeichen jeder Zeile spezifizieren den »Zeilentyp«. Beispielsweise ist der Typ der ersten Zeile oben `HEADER`, der der zweiten `TITLE`, später kommt dann ein `ATOM`.

Beispiele, wie Zeilentypen spezifiziert werden.

Für jeden Zeilentyp schreibt das Format dann genau vor, was danach kommen darf.

Beispiel (Der Zeilentyp `HEADER`)

Spalten	Datentyp	Beschreibung
1–6	Zeilentyp	Hier muss `RECORD` stehen
11–50	String	Klassifikation des Moleküls
51–59	Datum	Upload-Datum im Format DD-MMM-YY
63–67	Code	PDB-ID

Beispiel (Der Zeilentyp `TITLE`)

Spalten	Datentyp	Beschreibung
1–6	Zeilentyp	Hier muss `TITLE` stehen
9–10	Zahl	Laufende Nummer (mehrzeiligen Titel)
11–80	String	Der Titel des Experiments

Beispiele, wie Zeilentypen spezifiziert werden.

Beispiel (Der Zeilentyp `ATOM`)

Spalten	Datentyp	Beschreibung
1–6	Zeilentyp	Hier muss `ATOM` stehen
7–11	Zahl	Laufende Atomnummer (5′ beginnt).
13–16	Name	Atomname.
17	Zeichen	Indikator für alternativen Ort
18–20	String	Aminosäurenname
22	Zeichen	Ketten-ID
23–26	Zahl	Aminosäuren-Serienummer
27	Zeichen	Einfügecode für Säuren
31–38	Zahl	*x*-Koordinate in Ångström
39–46	Zahl	*y*-Koordinate in Ångström
47–54	Zahl	*z*-Koordinate in Ångström
55–60	Zahl	Occupancy
61–66	Zahl	Temperaturfaktor
77–78	String	Elementsymbol, rechtsbündig
79–80	String	Ladung

Die Biodatenbank Ensembl

Ensembl is a joint project between EMBL-EBI (European Molecular Biology Laboratory, The European Bioinformatics Institute) and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

- Die Biodatenbank Ensembl stellt Genomdaten ausgewählter Spezies zur Verfügung.
- Neben den Genomsequenzen sind Gendaten, deren Transkriptionen, Exons und vieles mehr gespeichert.

Zugriff ist auf viele Arten möglich:

1. Es gibt eine Webseite, auf der die Daten direkt durchsucht werden können.
2. Es werden auf der Webseite auch ein Reihe von speziellen Analyse-Tools angeboten.
3. Schließlich ist ein anonymer SQL-Zugang möglich.

Die Datenbank-Schemata von Ensembl.

Zentrale Schemata

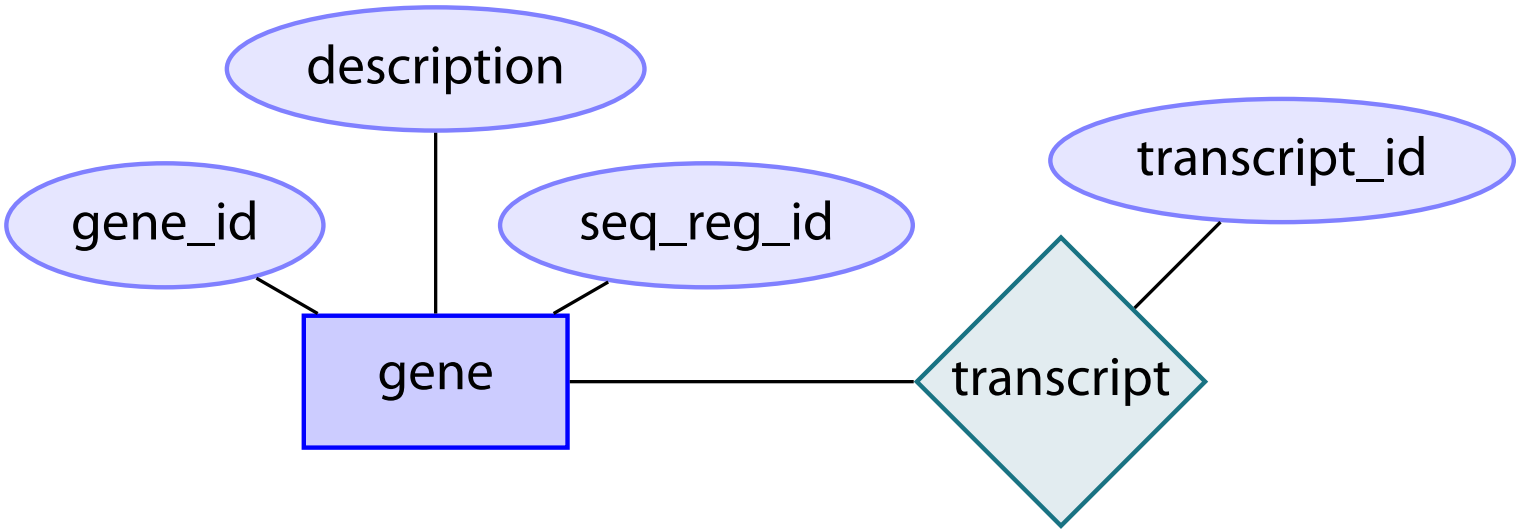
In Ensembl gibt es vier Hauptschemata:

- Der »Kern«, der Genomdaten und Gene speichert.
- »Vergleichende Genomik«, in der Homologe, Paraloge und Proteinfamilien gespeichert sind.
- »Variationen«, in der SNP-Varianten, somatische Mutationen und Strukturvarianten gespeichert sind.
- »Regulation«, in der regulatorische Motive gespeichert sind.

Viele Datenbanken

- In Ensembl gibt es pro Spezies mehrere Datenbanken.
- So wird immer, wenn genügend neue Daten gesammelt wurden, ein neuer »Build« erstellt und dieser in einer neuen Datenbank zur Verfügung gestellt.

Ein ganz kleiner Ausschnitt des Kernschemas



Ein typische Frage, die Ensembl beantworten kann.

Die zu beantwortende Frage

Welche für die *Phosporylasekinase* zuständigen Gene haben bei *Xenopus tropicalis* mehr als eine Transkription?

Schritt 1: Auswahl der richtigen Datenbank

```
> mysql --host=ensembldb.ensembl.org --user=anonymous
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 3872065
...
mysql> use xenopus_tropicalis_core_75_42;
```

Schritt 2: Erstellung der Anfrage

Zur Übung

Geben Sie SQL-Befehle an, die jeweils folgendes leisten:

1. Alle mit *Phosporylasekinase* befassten Gene auswählen.
2. Einen Join der Transkriptionstabelle mit der obigen Tabelle erstellen.
3. Die Anzahl der Transkriptionen pro Gen zählen, wobei die Ausgabespalten sein sollen: Anzahl der Transkriptionen, Gene-ID und Gen-Beschreibung.
4. Die Einschränkung der obigen Ausgabe auf Gene, die mindestens zwei Transkriptionen haben.

Die Anfrage


```
select    count(transcript_id) as c, gene_id, gene.description
     from transcript join gene using (gene_id)
    where gene.description like "%phosphorylase␣kinase%"
 group by gene_id, gene.description
   having count(transcript_id) > 1;
```

```
+---+---------+-----------------------------------------------+
| c | gene_id | description                                   |
+---+---------+-----------------------------------------------+
| 2 |    7298 | phosphorylase kinase, alpha 2 (liver) ...     |
| 2 |    8294 | calmodulin 1 (phosphorylase kinase, delta) ...|
+---+---------+-----------------------------------------------+
```

Wie man Ensembl »wirklich« nutzt.

- In der Praxis wird man nur sehr selten SQL-Anfragen »per Hand« erstellen.
- Vielmehr werden diese Anfragen *von Programmen* erstellt, die dann mit der Datenbank »sprechen«.
- Solche Programme kann man in Java schreiben und dann den *Java Database Connector* nutzen
- oder in Perl (eine Programmiersprache) und dort entsprechende Klassen und Methoden.

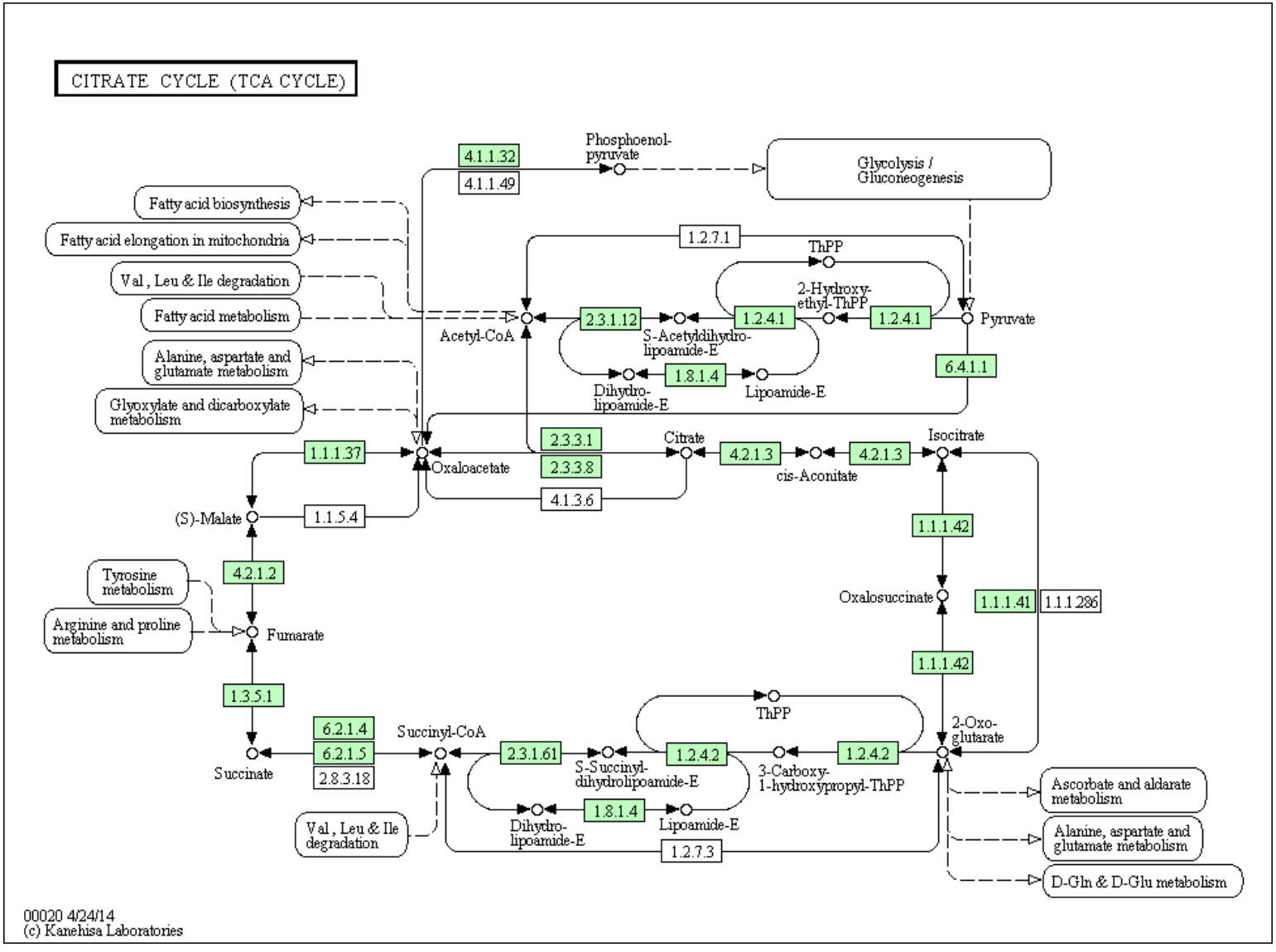
Die Biodatenbank Kyoto Encyclopedia of Genes and Genomes.

- Die KEGG-Biodatenbank speichert Gene und besonders *Pathways*.
- Die Pathways werden auch graphisch dargestellt und können durchsucht werden.

Zugriff ist wieder auf verschiedene Arten möglich:

1. Auf der Webseite kann man die Pathways anschauen und durch sie »mit der Maus« navigieren.
2. Die Pathways lassen sich auch als XML-Dateien herunterladen.

Der Citratzyklus als Bild



Der Citratzyklus als XML-Datei

Das KEGG stellt Pathways auch in der *KEGG Markup Language* KGML zur Verfügung:

```
<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM
     "http://www.kegg.jp/kegg/xml/KGML_v0.7.1_.dtd">
<pathway name="path:hsa00020" org="hsa" number="00020"
         title="Citrate␣cycle␣(TCA␣cycle)" ...>
    <entry id="28" name="ko:K00174␣ko:K00175␣ko:K00177␣ko:K00176"
         type="ortholog" reaction="rn:R01197" ...>
       <graphics name="K00174..." fgcolor="#000000"
              bgcolor="#FFFFFF" type="rectangle" x="526" y="649"
              width="46" height="17"/>
    </entry>
...
    <reaction id="92" name="rn:R00431␣rn:R00726"
         type="irreversible">
        <substrate id="60" name="cpd:C00036"/>
        <product id="94" name="cpd:C00074"/>
    </reaction>
</pathway>
```

Zusammenfassung

Was sind Biodatenbanken?

Eine *Biodatenbank* ist eine digitalisierte *Sammlung von biologischen Daten*.

In welchen Formaten stellen Biodatenbanken ihre Inhalte zur Verfügung?

- Flat-Files
- XML-Dateien
- SQL-Zugriff oder SQL-Dumps

Zusammenfassung

Was speichern Biodatenbanken?

- DNA-Sequenz-Datenbanken
- RNA-Sequenz-Datenbanken
- Protein-Sequenz-Datenbanken
- Molekülstruktur-Datenbanken
- Gen-Datenbanken
- Genexpressions-Datenbanken
- Pathway-Datenbanken
- Publikations-Datenbanken

Nucleotide Acids Research,
Database Summary Paper,
http://www.oxfordjournals.org/nar/database/cap,
Zugriff Juni 2014