Bio Data Management

Kapitel 3

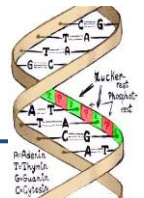
Datenmodelle und Anfragesprachen

Wintersemester 2014/15
Dr. Anika Groß
Universität Leipzig, Institut für Informatik, Abteilung Datenbanken
http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

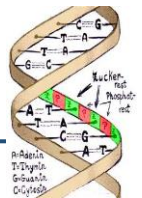
Vorläufiges Inhaltsverzeichnis

1. Motivation und Grundlagen
2. Bio-Datenbanken
3. Datenmodelle und Anfragesprachen
4. Modellierung von Bio-Datenbanken
5. Sequenzierung und Alignments
6. Genexpressionsanalyse
7. Annotationen
8. Matching
9. Datenintegration: Ansätze und Systeme
10. Versionierung von Datenbeständen
11. Neue Ansätze



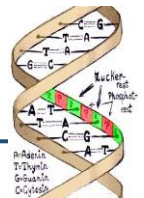
Lernziele

- Kennenlernen und Wiedergabe
  - Verschiedener Datenmodelle und Anfragesprachen, die in den Lebenswissenschaften Verwendung finden, sowie
  - deren Vor- / Nachteile



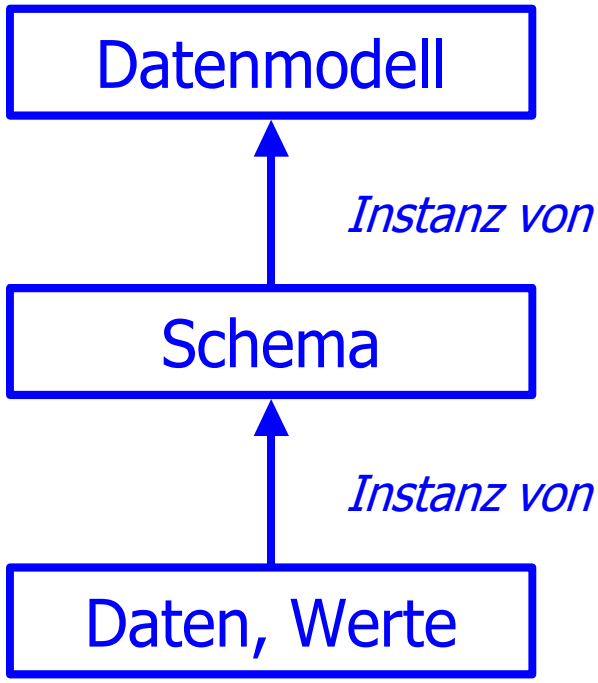
Gliederung

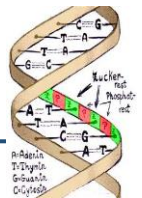
- Grundbegriffe
  - Daten, Schema, Datenmodell
  - Granularität von Datenmodellen
- Datenmodelle
  - Entry-basiertes Datenmodell
  - ASN.1
  - Relationales Modell
  - Objektorientierte Modelle
  - XML-basierte Modelle



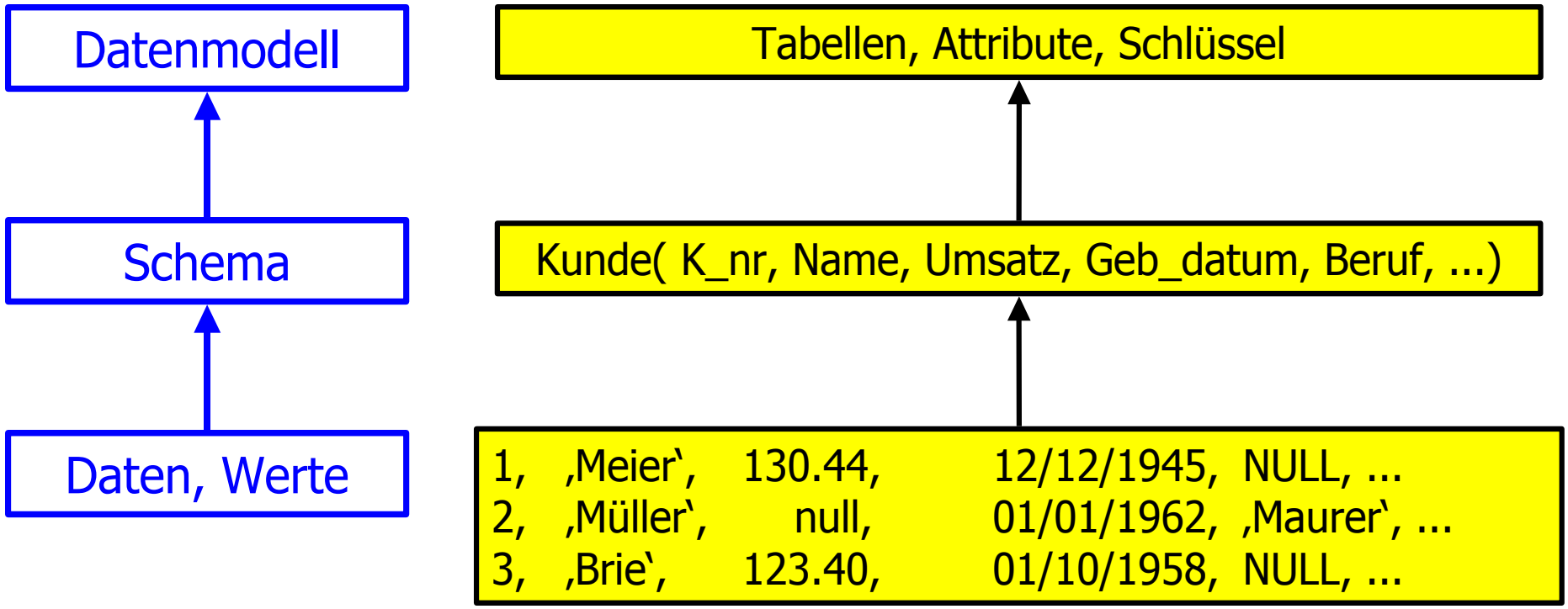
Grundbegriffe

- Daten
  - Tatsächliche Werte, uninterpretiert
  - Ergebnisse von Anfragen
- Schema
  - Beschreibt Typ und Organisation der Werte
  - Spezifiziert durch DDL (z.B. SQL)
- Datenmodell
  - Definition der Modellierungsprimitive, aus welchen ein Schema bestehen kann
  - RDBMS: Tabellen, Attribute, …
  - ORDBMS: Klassen (UDTs), Attribute, Methoden, …



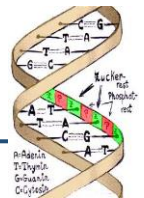


Beispiel: RDBMS



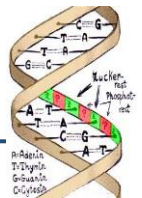
Granularität eines Datenschema

- "Breite Datenbanken"
  - "Wenig Klassen (wenig Detailtiefe), viele Objekte" (d.h. von vielen Objekten wird relativ wenig Information gespeichert)
  - EMBL-Sequenzdatenbank, ArrayExpress, 2D Page, …
- "Tiefe Datenbanken"
  - "Viele Klassen (große Detailtiefe), wenig Objekte" (d.h. von wenigen Objekten wird relativ viel Information gespeichert)
  - Chromosom- / Spezies- / Krankheitsspezifische Datenbanken



Austauschformate

- Verschiedene Austauschformate
  - FASTA, EMBL Format
  - ASN.1 (Sequenzen + Annotationen)
  - MAGE (Experimentannotation bei Expressionsexperimenten)
- Export üblicherweise in Flat Files, XML
- Relationale Systeme
- XML DTD's definiert für verschiedene Projekte, z.B.
  - GAME: Genome Annotation Markup Elements
  - BIOML: BIOpolymer Markup Language
  - BSML: Bioinformatic Sequence Markup Language



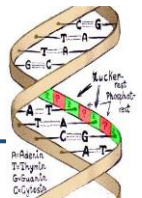
FASTA-Format

- Textbasiertes Format zur Darstellung und Speicherung von Sequenzen
- Mehrere Sequenzen pro Datei möglich
- Kopfzeile einer Sequenz zur Beschreibung: beginnt mit „>“, danach mehrere IDs und Namen mit „|“ getrennt
  - z.B. von Genbank: gi|*gi-number*|gb|*accession*|*locus*
- Kommentarzeile startet mit „;“
- Sequenzdarstellung: ein oder mehrere Zeilen (ca. 80 Zeichen pro Zeile), Protein- oder Nukleinsäuresequenzen (Ein-Buchstaben-Code), Lücken und Alignierungszeichen erlaubt
- Beispiel: Proteinsequenz im FASTA-Format vom Cytochrom b des Asiatischen Elefanten (http://de.wikipedia.org/wiki/FASTA-Format):

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

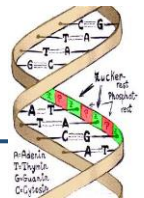
Datenmodelle und -systeme

- Entry-basiertes Datenmodell
- ASN.1
- Relationales Modell
- Objektorientierte Modelle
- XML-basierte Modelle



Entry-basiertes Datenmodell

- Kein Datenmodell im eigentlichen Sinn (wie z.B. RM, OO)
- Flat-file
- Weite Verbreitung in Life Sciences
  - EMBL, Swiss-Prot, Interpro, Omim, Genbank ,…
  - Unterstützt von vielen Bio-Datenbanken
- Beispiel Swiss-Prot
  - Menge von Proteinsequenzen
  - Core-Elemente: Sequenz, Taxonomie, Zitierung
  - Annotationen: Domainen, Sequenzvarianten, assoziierte Krankheiten, Sekundärstruktur, …



Entry-Modell

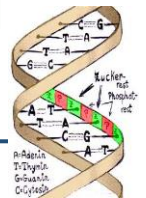
```
ID   INS_HUMAN               Reviewed;         110 AA.
AC   P01308; Q5EEX2;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   21-JUL-1986, sequence version 1.
DT   21-MAR-2012, entry version 168.
DE   RecName: Full=Insulin;
DE   Contains:
DE     RecName: Full=Insulin B chain;
DE   Contains:
DE     RecName: Full=Insulin A chain;
DE   Flags: Precursor;
GN   Name=INS;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=80120725; PubMed=6243748; DOI=10.1038/284026a0
RA   Bell G.I., Pictet R.L., Rutter W.J., Cordell B., Tischer E.,
RA   Goodman H.M.;
RT   "Sequence of the human insulin gene.";
RL
```

Microsyntax,
Feldabhängige Formate, NF²

Unkontrollierte
Vokabulare

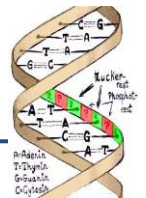
Eingebettete Objekte
(keine Verweise)

Line codes (pre-XML): Referenz auf (Record-)Struktur einer Zeile
AC=Accession Code, DE = Description, DT = Date ,OS = Organism, OC =Taxonomy



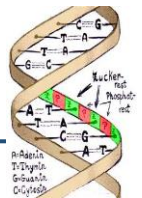
Line Codes

Line code	Content	Occurrence in an entry
ID	Identification	Once; starts the entry
AC	Accession number(s)	One or more
DT	Date	Three times
DE	Description	One or more
GN	Gene name(s)	Optional
OS	Organism species	One or more
OG	Organelle	Optional
OC	Organism classification	One or more
RN	Reference number	One or more
RP	Reference position	One or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RA	Reference authors	One or more
RT	Reference title	Optional
RL	Reference location	One or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
KW	Keywords	Optional
FT	Feature table data	Optional
SQ	Sequence header	Once
	(blanks) sequence data	One or more
//	Termination line	Once; ends the entry

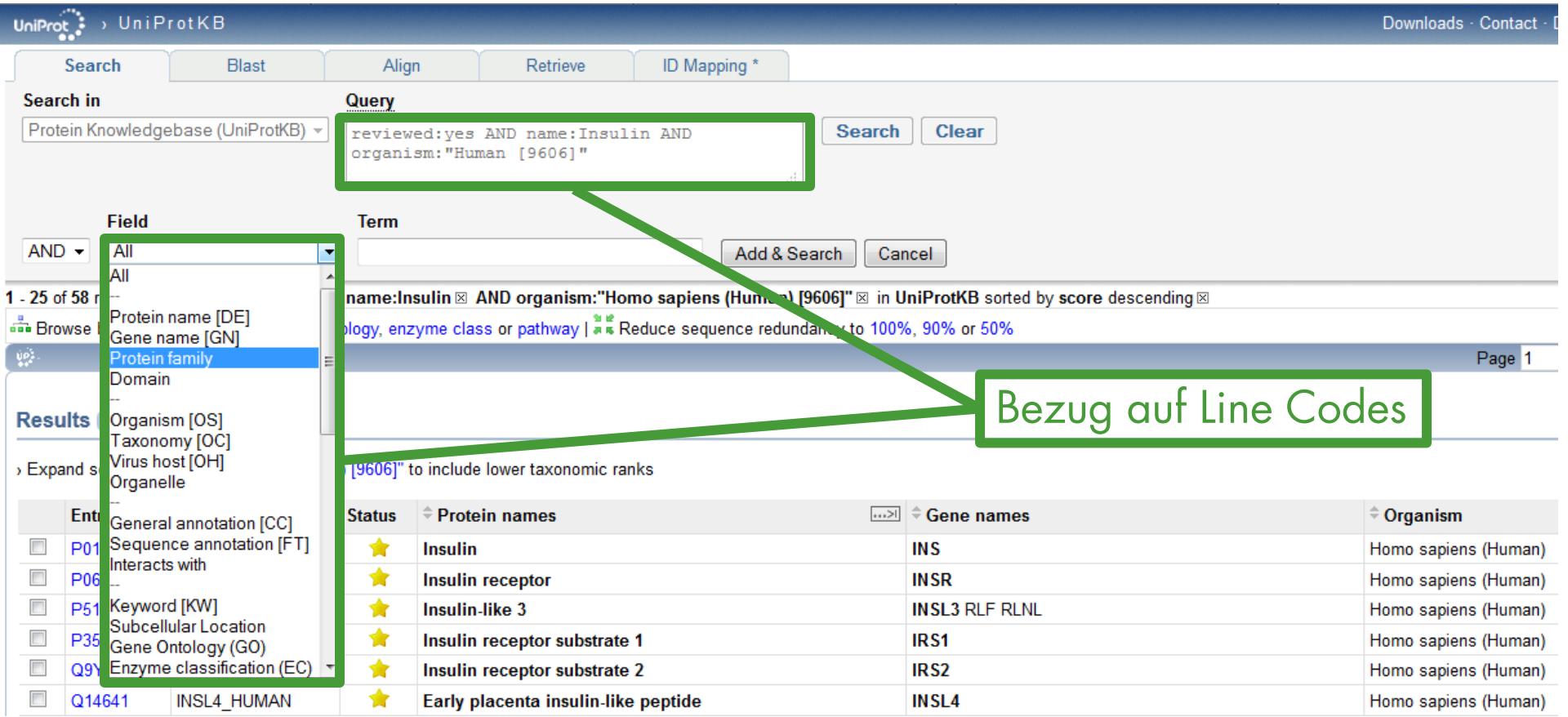


Entries

- Datenbank ≙ Menge ähnlich strukturierter Entries
- Entry: Menge von Feldern (Attribute, Lines) zu einem Bio-Objekt (z.B. zu einem Protein)
  - Identifikation durch (standardisierten) Line Code
  - Können 0-n mal vorkommen (semistrukturiert)
  - Können komplexe eigene Struktur haben
  - Können eingebettete Objekte repräsentieren
  - Microsyntax in Werten (Sprechende Schlüssel)
- Keine deklarativen Konsistenzbedingungen
- Kein Klassen- oder Objektbegriff



Entry-Modell: Anfrage (Swiss-Prot)



Names and origin

Protein names	*Recommended name:* **Insulin** Cleaved into the following 2 chain 1. ***Insulin B chain*** 2. ***Insulin A chain***
Gene names	Name:**INS**
Organism	**Homo sapiens (Human)** [Reference proteome]
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertel

Web-Darstellung: http://www.uniprot.org/uniprot/P01308
txt: http://www.uniprot.org/uniprot/P01308.txt

Protein attributes

Sequence length	110 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a
Protein existence	Evidence at protein level

General annotation (Comments)

Function	Insulin decreases blood glucose concentration. It i glycogen synthesis in liver.
Subunit structure	Heterodimer of a B chain and an A chain linked by
Subcellular location	Secreted.
Involvement in disease	Defects in INS are the cause of familial hyperproins Defects in INS are a cause of diabetes mellitus ins susceptibility to ketoacidosis in the absence of ins secondary thirst. These derangements result in lor Defects in INS are a cause of diabetes mellitus pe type 1. It is characterized by insulin-requiring hype Defects in INS are a cause of maturity-onset diabe of inheritance, onset in childhood or early adulthoo Ref.33 Ref.34 Ref.35
Pharmaceutical use	Available under the names Humulin or Humalog (E Lys-53.
Sequence similarities	Belongs to the insulin family.

Ontologies

Keywords

Biological process	Carbohydrate metabolism Glucose metabolism
Cellular component	Secreted
Coding sequence diversity	Polymorphism
Disease	Diabetes mellitus Disease mutation

```
ID   INS_HUMAN               Reviewed;         110 AA.
AC   P01308; Q5EEX2;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   21-JUL-1986, sequence version 1.
DT   06-MAR-2013, entry version 178.
DE   RecName: Full=Insulin;
DE   Contains:
DE     RecName: Full=Insulin B chain;
DE   Contains:
DE     RecName: Full=Insulin A chain;
DE   Flags: Precursor;
GN   Name=INS;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
...  ... ... ... ... ... ... ... ... ... ...
DR   Pathway_Interaction_DB; insulin_pathway; Insulin Pathway.
DR   Pathway_Interaction_DB; insulin_glucose_pathway; Insulin-mediated glucose transport.
DR   Pathway_Interaction_DB; mtor_4pathway; mTOR signaling pathway.
DR   Pathway_Interaction_DB; ptp1bpathway; Signaling events mediated by PTP1B.
DR   Reactome; REACT_111045; Developmental Biology.
DR   Reactome; REACT_111102; Signal Transduction.
DR   Reactome; REACT_111217; Metabolism.
DR   Reactome; REACT_116125; Disease.
DR   ChEMBL; CHEMBL5881; -.
DR   EvolutionaryTrace; P01308; -.
DR   GenomeRNAi; 3630; -.
DR   NextBio; 14203; -.
DR   PMAP-CutDB; P01308; -.
DR   ArrayExpress; P01308; -.
DR   Bgee; P01308; -.
DR   Genevestigator; P01308; -.
DR   GermOnline; ENSG00000129965; Homo sapiens.
DR   GO; GO:0005788; C:endoplasmic reticulum lumen; TAS:Reactome.
DR   GO; GO:0031904; C:endosome lumen; TAS:Reactome.
DR   GO; GO:0005615; C:extracellular space; IDA:BHF-UCL.
DR   GO; GO:0005796; C:Golgi lumen; TAS:Reactome.
DR   GO; GO:0030141; C:secretory granule; TAS:Reactome.
DR   GO; GO:0005179; F:hormone activity; NAS:UniProtKB.
DR   GO; GO:0005158; F:insulin receptor binding; IDA:UniProtKB.
```

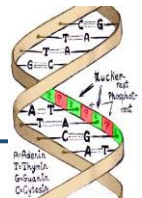
Entry Modell

→

Relationales Modell

```
ID   INS_HUMAN               Reviewed;         110 AA.
AC   P01308; Q5EEX2;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   06-MAR-2013, entry version 178.
DE   RecName: Full=Insulin;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria;
OC   Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
SQ   SEQUENCE   110 AA;  11981 MW;  C2C3B23B85E520E5 CRC64;
     MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY TPKTRREAED
     LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC SLYQLENYCN
RL   BMC Med. Genet. 11:42-42(2010).
CC   -!- FUNCTION: Insulin decreases blood glucose concentration. It increases cell permeability
CC       to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose
CC       phosphate cycle, and glycogen synthesis in liver.
CC   -!- DISEASE: Defects in INS are the cause of familial hyperproinsulinemia (FHPRI) [MIM:176730].
DR   Ensembl; ENST00000250971.
DR   Ensembl; ENST00000381330.
DR   UCSC; uc001lvn.2; human.
DR   Reactome; REACT_111217; Metabolism.
DR   Reactome; REACT_116125; Disease.
DR   GO; GO:0005615; C:extracellular space; IDA:BHF-UCL.
DR   GO; GO:0005179; F:hormone activity; NAS:UniProtKB.
DR   GO; GO:0015758; P:glucose transport; IDA:UniProtKB.
DR   GO; GO:0008286; P:insulin receptor signaling pathway; TAS:Reactome.
DR   InterPro; IPR004825; Insulin.
DR   PRINTS; PR00277; INSULIN.
DR   PRINTS; PR00276; INSULINFAMLY.
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=80120725; PubMed=6243748; DOI=10.1038/284026a0;
RA   Bell G.I., Pictet R.L., Rutter W.J., Cordell B., Tischer E., Goodman H.M.;
RT   "Sequence of the human insulin gene.";
RL   Nature 284:26-32(1980).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RX   MEDLINE=80236313; PubMed=6248962; DOI=10.1126/science.6248962;
RA   Ullrich A., Dull T.J., Gray A., Brosius J., Sures I.;
RT   "Genetic variation in the human insulin gene.";
RL   Science 209:612-615(1980).
//
```

Bsp. Auszug aus Swiss-Prot
Entry INS_HUMAN



Entry				
ID (PK)	TaxID (FK)	Name	Sequence CRC64	AA Sequence
INS_HUMAN	9606	Insulin	C2C3B23B..	MALWMRLLPL...

Accession Number	
ID (FK)	Acc (PK)
INS_HUMAN	P01308
INS_HUMAN	Q5EEX2

Species	
TaxID (PK)	Name
9606	Homo Sapiens

Entry Dates		
ID (PK,FK)	Date (PK)	Reason (PK)
INS_HUMAN	21-JUL-1986	integrated into UniProtKB/Swiss-Prot
INS_HUMAN	06-MAR-2013	entry version 178.

EC = Evidence code:
IDA Inferred from direct assay
NAS Non-traceable Author Stat.
TAS Traceable Author Statement

External Object Reference		
ID (PK,FK)	Acc (PK,FK)	EC
INS_HUMAN	ENST00000250971	
INS_HUMAN	ENST00000381330	
INS_HUMAN	GO:0005615	IDA
INS_HUMAN	GO:0005179	NAS
INS_HUMAN	GO:0015758	IDA
INS_HUMAN	IPR004825	TAS

External Object		
Acc (PK)	Name	Data source (FK)
ENST00000250971		Ensembl
ENST00000381330		Ensembl
GO:0005615	extracellular space	GO
GO:0005179	hormone activity	GO
GO:0015758	glucose transport	GO
IPR004825	Insulin	InterPro

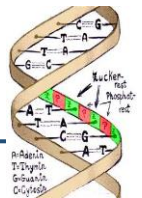
Data Source
Name (PK)
InterPro
GO
Ensembl

Literature Reference	
ID (PK,FK)	Pubmed Acc (PK,FK)
INS_HUMAN	6243748
INS_HUMAN	6248962

Article			
ID (PK,FK)	Title	Year	Journal/Venue
6243748	Sequence of the human insulin gene	1980	Nature
6248962	Genetic variation in the human insulin gene	1980	Science

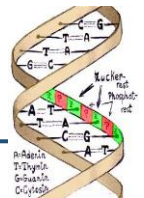
Author		
ID (PK,FK)	Vorname	Nachname
1	B.G.	Bell
2	R.L.	Pictet
3	A.	Ullrich
4	T.J.	Dull

Author_Publication	
AutorID (PK,FK)	PubID (PK,FK)
1	6243748
2	6243748
3	6248962
4	6248962



Entry-Modell: Zusammenfassung

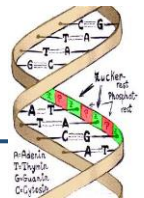
- Datenmodell
  - Einziges Modellierungsprimitiv: Der Entry
  - Felder mit hierarchischer Schachtelung; Keine Assoziationen
- Schema: Keine explizite Repräsentation auf Basis einer DDL
- Werte: Einfache und zusammengesetzte Werte möglich
- Eher Format als Datenmodell
- Vorteile
  - Sofort lesbar für Menschen, plattformunabhängig (ASCII), hohe Flexibilität durch textorientieres Editieren, leicht zu durchsuchen (Grep, "search"-Button)
- Nachteile
  - Keine Konsistenzbedingungen
  - Hohe Redundanz durch viele geschachtelte Objekte: Literatur, Taxonomie, Cross-Referenzen…
  - Keine strukturierten Anfragen möglich



ASN.1

- Abstract Syntax Notation One (http://asn1.elibel.tm.fr/)
- Zur abstrakten Beschreibung von Datentypen
- Zusammenfassung zu Modulen
- Schnittstelle des Moduls: IMPORT und EXPORT
- Formatbeschreibungssprache ähnlich DTD / XML Schema
- Ursprünglich für Definition von Datenaustauschformaten in der Telekommunikationsbranche
- Internationaler Standard (1984) ISO 8824 / 8825
- Verwendet von NCBI*-Datenbanken
  - Genbank
  - UniGene
  - dbSNP
  - …

* National Center for Biotechnology Information

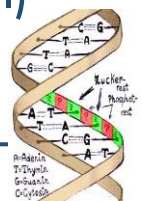


ASN.1: Elemente

- Datenmodell mit expliziten Typen (im Gegensatz z.B. zum Entry-Modell)
- BNF-ähnlich
- Ein Typ besteht aus
  - Primitiven Attributen
  - Strukturen (structs)
  - Mengenwertige Attribute meist durch spezielle Kodierung
  - Choices (Varianten)

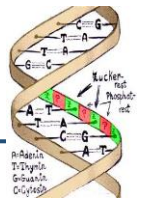
```
Married ::= BOOLEAN
Age ::= INTEGER
Picture ::= BIT STRING
Form ::= SEQUENCE { name                 PrintableString,
                    age                  Age,
                    married              Married,
                    marriage-certificate Picture OPTIONAL }

Quantity ::= CHOICE { units       INTEGER,
                      millimeters INTEGER,
                      milligrams  INTEGER }
```

- Definition von Modulen: Sammlung von Typdefinitionen und/oder Wertedefitionen
- Elementare Datentypen: `INTEGER, BOOLEAN, CHARACTER STRING …`
- Zusammengesetzte Datentypen: `SEQUENCE, SET, …`
- Constraints, z.B. `Lottery-number ::= INTEGER (1..49)`
- Language Mappings verfügbar (für Java, C++, C, COBOL, XML, Perl)
- XML-Mapping von NCBI-Datenbanken verwendet (ASN2XML)



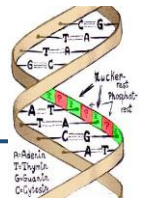
Moduldefinition

```
Demo-module DEFINITIONS ::=          -- Modulname DEFINITIONS ::= BEGIN
BEGIN

EXPORTS My-type;                     -- My-type kann von and. Modulen
                                        genutzt werden
IMPORTS Foreign-type FROM Other-module;  -- Importieren von Typen

My-type ::= SEQUENCE {               -- Definieren ein Objekt “My-type”
                                     -- My-type ist eine Typreferenz
   first     INTEGER ,               -- first ist ein Identifier,
   second    INTEGER DEFAULT 2 ,     -- second ist per default 2
   third     VisibleString OPTIONAL -- third ist ein optionaler String

}                                    -- Ende der Objektdefinition
Another ::= Foreign-type             -- andere definierte Typen können
                                        referenziert werden
END                                  -- Ende des Moduls (END required)
```



Werte / Daten

```
My-type ::= {                          -- first = 42, second = 2,
    first 42                            third existiert nicht
}

My-type-set ::= SET OF My-type         -- mehr als einen Datensatz

My-type-set ::= {                      -- starte SET OF
   {                                   -- 1. My-type
      first 42
   } ,
   {                                   -- 2. My-type
      first 27 ,
      second 22 ,
      third "Everything set here"
   }
}
```



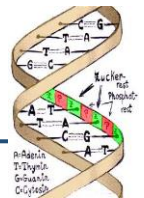
ASN.1-Beispiel: NCBI-PubMed Modul

```
NCBI-PubMed DEFINITIONS ::=
BEGIN

EXPORTS Pubmed-entry, Pubmed-url;
IMPORTS PubMedId FROM NCBI-Biblio
        Medline-entry FROM NCBI-Medline;

Pubmed-entry ::= SEQUENCE {          -- a PubMed entry
    -- PUBMED records must include the PubMedId
    pmid PubMedId,
    -- Medline entry information
    medent Medline-entry OPTIONAL,
    -- Publisher name
    publisher VisibleString OPTIONAL,
    -- List of URL to publisher cite
    urls SET OF Pubmed-url OPTIONAL,
    -- Publisher's article identifier
    pubid VisibleString OPTIONAL
}

Pubmed-url ::= SEQUENCE {
    location VisibleString OPTIONAL, -- Location code
    url VisibleString                -- Selected URL for location
}

END
```

Modularer Aufbau

Export / Import von Modulen

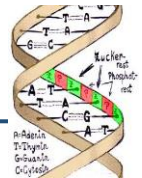
Primitive / komplexe Datentypen



ASN.1-Beispiel: NCBI Sequenzeintrag

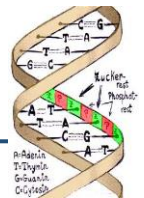
```
Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  release "" ,
  descr {
    source {
      org {
        taxname "Adeno-associated virus 2" ,
        db {
          {
            db "taxon" ,
            tag
              id 10804 } } ,
        orgname {
          name
            virus "Adeno-associated virus 2" ,
          lineage "Viruses; ssDNA viruses;
                   Parvoviridae;
                   Parvovirinae;
 Dependovirus" ,
          gcode 1 ,
          inst {
            repr raw ,
            mol dna ,
            length 4675 ,
            strand ss ,
            seq-data
              ncbi2na 'FA51D5DDE6676767478A5A98502B656195A9FE56A6974B898989999222A2
E941D4D1CAAF5E8AAE8B6E1B83C6D32AF2A2B5ECF22B46E2EFF984FF98453BAD19EACFC2562E24
64AB753FE09A8AFE0664965396ABFC623EE3C2B552617E1A93795A4FDE127FB81EAE9620A0EA2F
96523DE13A378378F892915785BA5882792661FDE1A0E996EE2C29568A57FDFEE43F882A22271F
51391B9DBA0145AAE0353AFFA86FD78B48F66001E3D220FC59A8D8961FE501EBD9AD1021480E99
68A6A042BAE8E2E71355OF1F9D5401525E2752EA6E870CE812CFC265EFE0DD1A26C06BEBA64937
8646ED92192892040883483543DE3996B8D234007D252B13A27ADABA76E842A8F1768824BA3528
A14A5D31375F439A5D41D9AD503429E5FA10E6A023CE25E1C0165561C5EBA949256E8A13F52436
8FCC03FE81C06AC63550CE69F5B7F7A8EA51800BDA4228114DE9EFEA5E41C5A821413668A53254
47B95F71AB9B01E850E207F57D061EED8423AE37AEA28AA08E1650ADBA2DA50253DDA28242B99B
A1480E42D7695232158756E36D17504504EE65B8F86A074185F60452496F9085A3BD03F81D1596
DE8D387FA82B4509282D021FFD6BA9028D1BAF8AE89383DC6D002AE894200856554B8648CC2E25
406AE662D2F9925361B4866827D8D0719212B140103BDDB46EA4E0DE39EFD5E4843988838 3483D
03379F4746848087BF22E7F56ED20DD056FDEDB400A6CD201EE713D34CD3A80AE5219F91E5E637
ADOEE8FE8E1E4DFE043038FC0D2B3A7963AF37D48FA76284777778283084BAE827405E951451409
5922693OA18492AB7EE7D7AB10B176857D0687610A8896B418A48659A576246C4097185A49D849
A21056C5D0B10519619A2FD28997C08231B7FEAA41768624B7D4A60022AF7E05DEA5EBE2817BC2
1A75A8002296B224775EE894875D76A05A029A94925E420023E0FFAD21E8864874B17855497768
494524955DEB7A81C318E9C4A4BA6450E9210C18A9961A2EAC3D75A03E93E63D44E8EA6122D345
14915817A95E545C410517710103F5250D289760610D1C7FA712455FAAB3FE1F4123D47947FD14
6E1E9021D341041EA8F5854221D07D09DFC13D0B408AD1920E1AC61863E50C17C5246BD2BBF1E1
DA2C52756C6D769DA64D0A397565BD52486DF4EB944B3A3174578106A2D292CA19DF4FF1E5E8B1
FD7DD239E6C5A0107F17D2711FF8A1BD7F51249C6744948B7A16DD3835DD3614B17B3C7E248101
1D42E8145192D0A7D2FF74A5A262E13DA852DCA07A7D7A15EF1652498B34084DE68C10412E0C76
E87A27142C51743A488777AE0D694E90944286382000BFFD7489AAF74DFEA090A74880103B813E0
0AD38F12182280DA0414356E9C6892CEBDECDC505D488A4121092716 48ED044429BDF5293ADE92
8488EEC5F4AA54DEA408F5111A1A13FD157755D3ABA3DA1F0115D75123DD34204 55AC5E60D7D85
17D2E6902FE7D7D3444B1D46A11AD26E88D8BA89E48280124067A0D580F4B11F5071042DEF0DBA
1F16E8C70E9BB3D225D954FA452317876C37B0F9FBC343016FC3DBF4BE07FADDE6CFDFDF372FD4
E9C6C8C2C93A6AF0D3C1C4281572E3A2FA51D5DDE6676767478A5A98502B656195A9FE56A6974B
898989999222A2E940'H } ,
          annot {
            {
              data
```

http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDKDOCS/DATAMODL.HTML



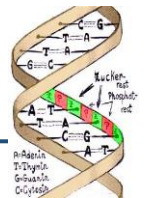
ASN.1: Zusammenfassung

- Vorteile
  - Binäres Encoding sehr kompakt (1 Base – 2 Bit)
  - Vollständige Toolbox erhältlich (NCBI)
  - Plattformunabhängig
- Nachteile
  - Binäres encoding für Menschen unlesbar
  - Text Encoding schwierig zu parsen
- Zukunft in Life Science unklar
- Vermutlich Ablösung durch XML

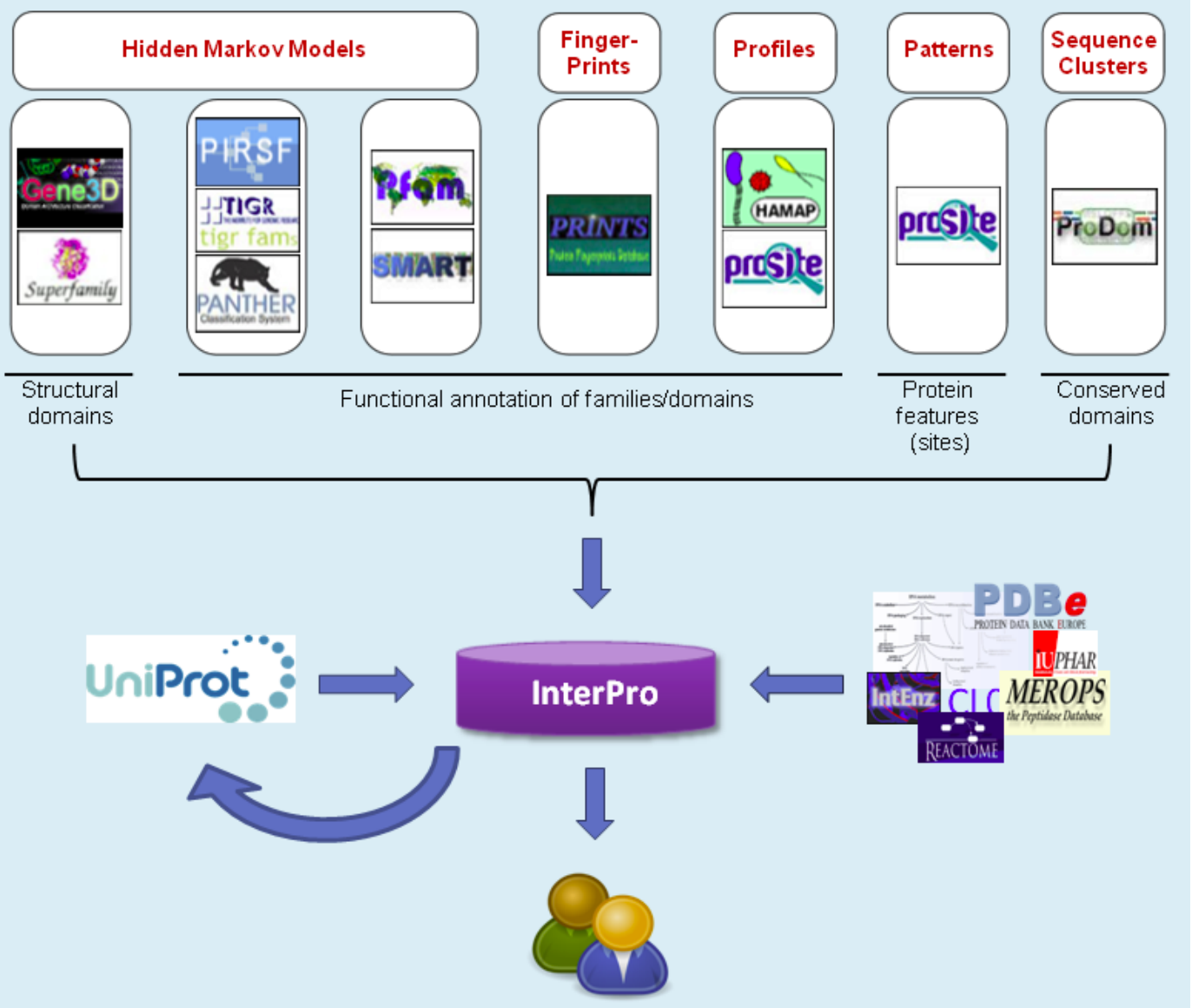


Relationales Modell

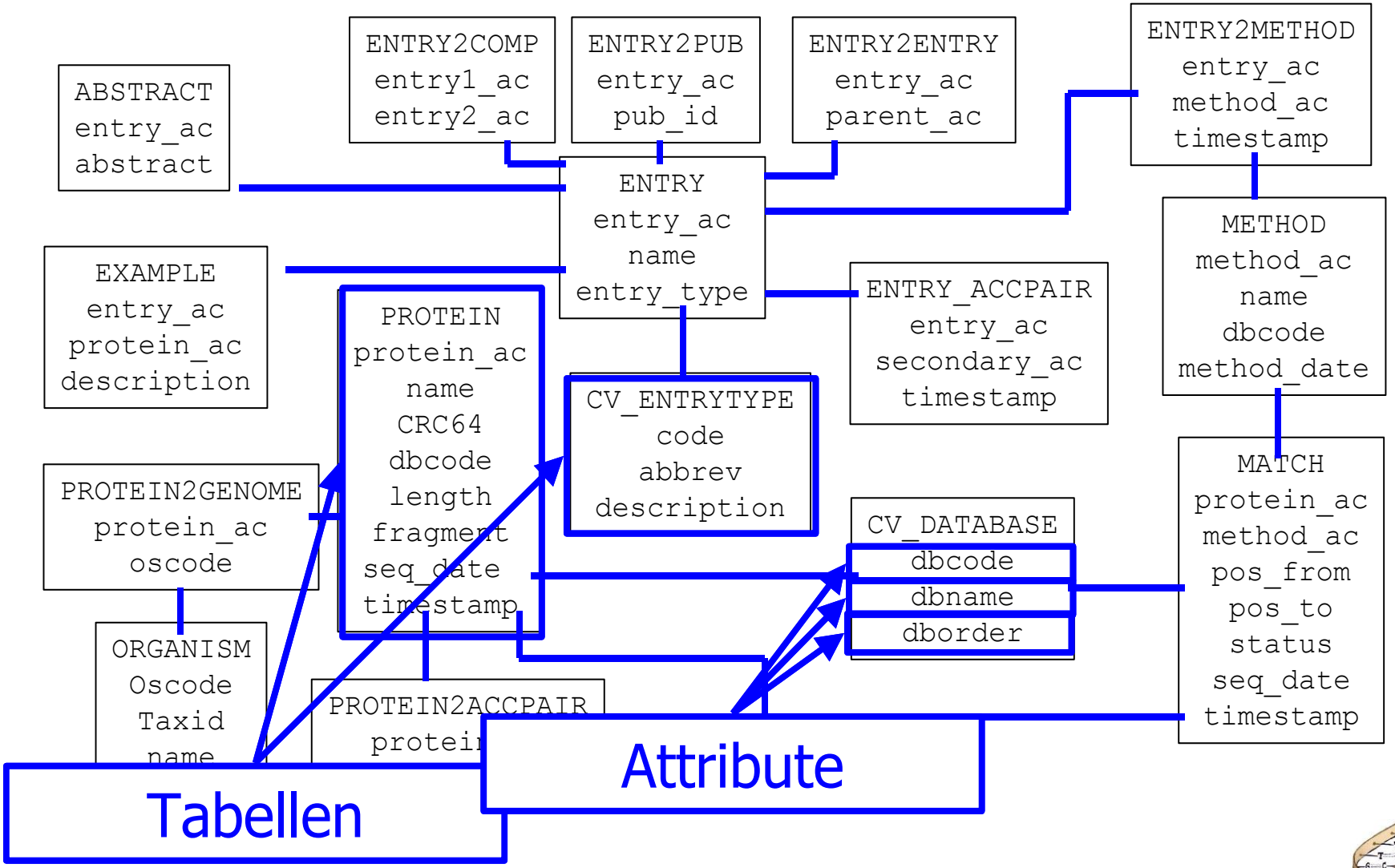
- Industriestandard
- Konzentration auf Speicherung/Retrieval
  - Semantisch arm, wenig Elemente
  - Nicht als Designmodell gedacht (wie z.B. ER oder UML)
  - User-Interfaces müssen programmiert werden
- Entwickelt für Transaction-Processing, Mehrbenutzerbetrieb, Client-Server
  - Overhead für typische "Read-Only" Bio-DB's
  - Komplizierte Installation, Administration, Backup, …



Beispiel InterPro

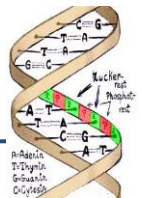


InterPro - Relationales Schema (Auszug)

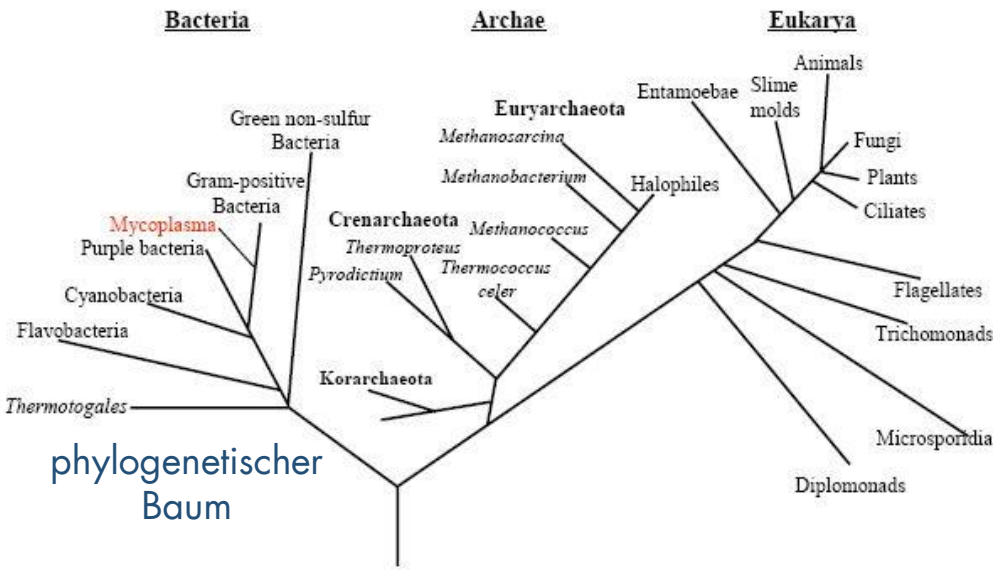


Bio* Projekte

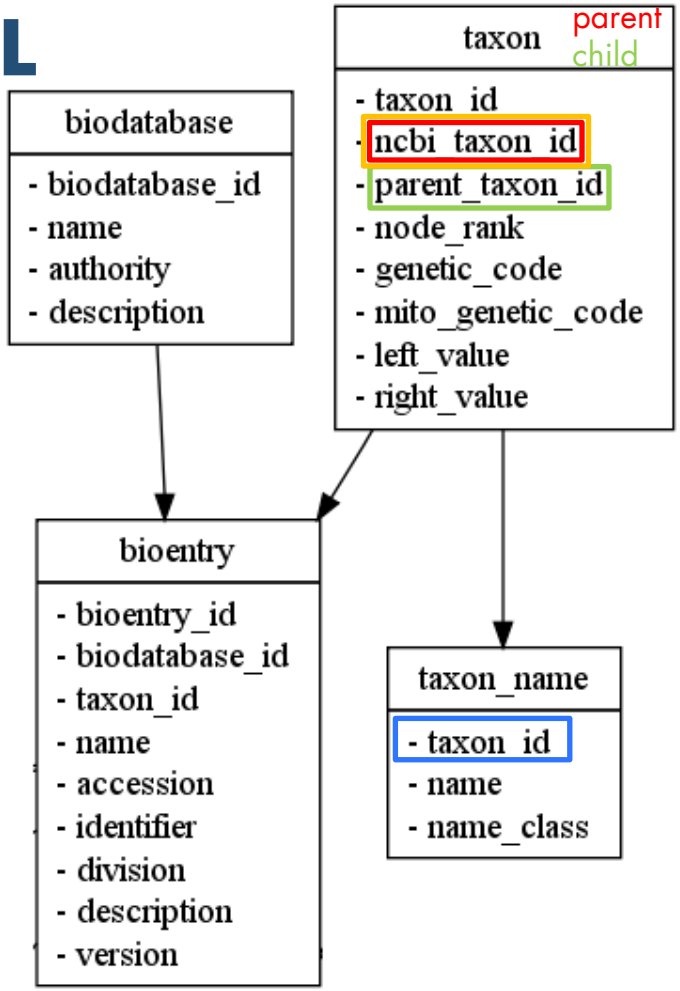
- BioSQL (http://www.biosql.org) als Grundlage für verschiedene Softwareprojekte: BioJava, BioPerl, BioPython, …
- Universelles relationales Schema zur Verwaltung und Analyse von biologischen Objekten
  - Anlehnung an Entry: BioEntry, Sequenzdaten und Beziehungen zwischen den Objekten
  - Referenzen zu weiteren Datenquellen
  - Verwendung von Ontologien/Taxonomien für Beschreibung
  - Implementierungen für verschiedene RDBMS (PostgreSQL, MySQL, Oracle, HSQLDB, and Apache Derby)
  - http://www.biosql.org/wiki/Schema_Overview
- Modularisierter Aufbau: Core Schema + Extension Modules, z.B. zur Verwaltung von phylogenetischen Bäumen



Spezies Taxonomie in BioSQL

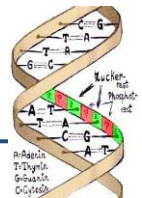


phylogenetischer Baum



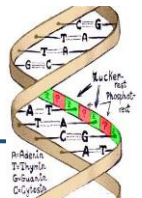
Finde die Taxon ID des Eltern-Taxon für 'Homo sapiens' (self-join):

```
SELECT parent.ncbi_taxon_id
FROM taxon AS parent JOIN taxon AS child
ON child.parent_taxon_id = parent.ncbi_taxon_id
JOIN taxon_name ON taxon_name.taxon_id = child.ncbi_taxon_id
WHERE taxon_name.name = 'Homo sapiens';
```



Relationales Modell: Zusammenfassung

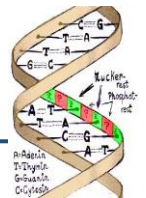
- Datenspeicherung und -retrieval
- Vorteile
  - Strukturierte Anfragen
  - Sehr weit verbreitet, robust, Industriestandard
  - Skalierbarkeit und Optimierbarkeit
  - Viele Produkte verfügbar
  - Ständige Weiterentwicklung
- Nachteile
  - SQL schwierig zu lernen
  - Volltextsuche nicht direkt möglich
  - Datenaustausch zw. Forschungsgruppen schwieriger als mit Entry-basierten Flatfiles



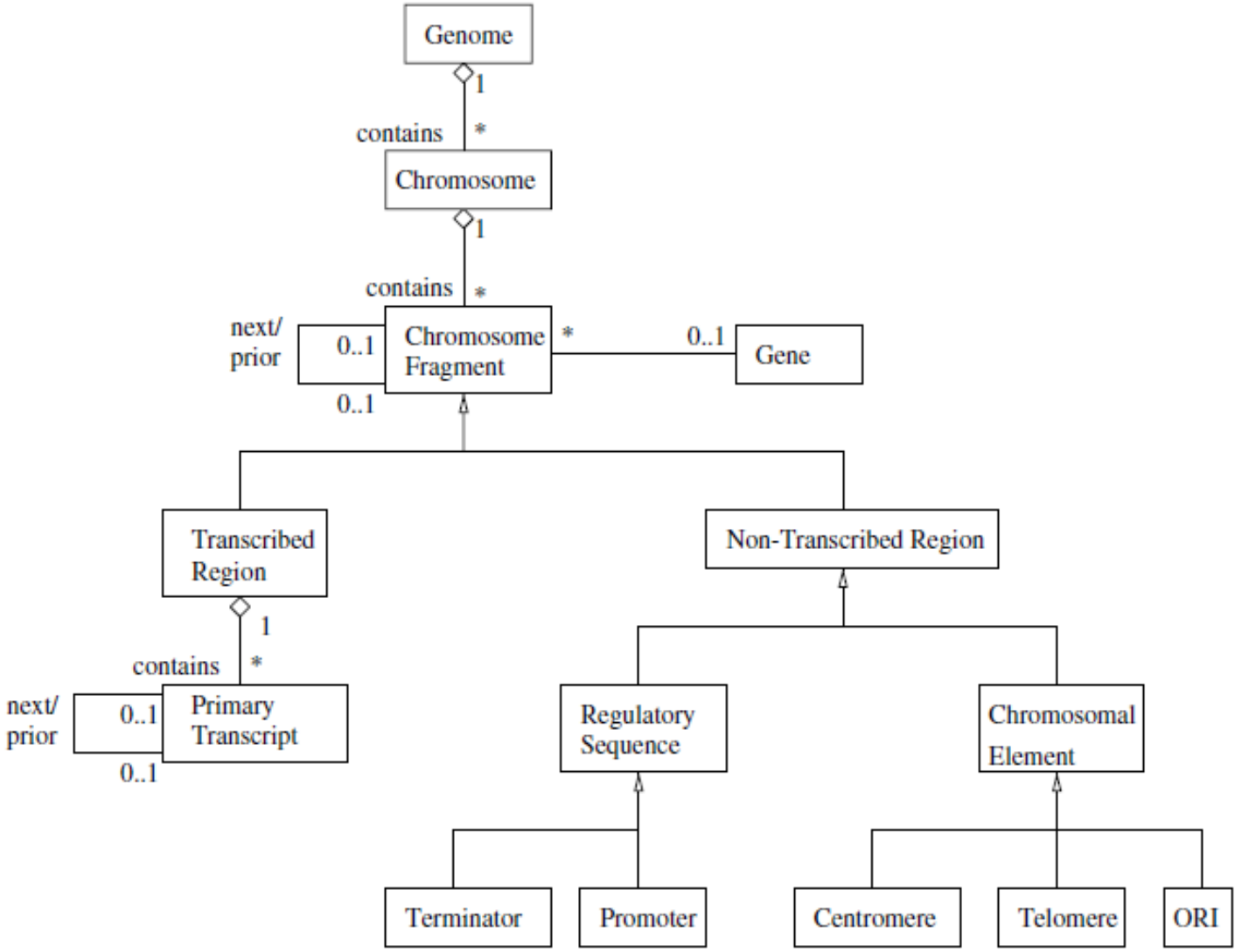
Objektorientiertes Modell

- UML (Unified Modelling Language), OPM (Object-Protocol Model), ACeDB
- UML: Industriestandard zur Modellierung von
  - Software: Klassen-Diagramme, Sequenzcharts, Zustands-Diagramme
  - Architekturen: Verteilungs-Diagramme, Komponenten-Diagramme
  - Requirements: Use Cases
  - Prozessen: Aktivitäts-Diagramme, Collaboration-Diagramme
- Viele UML-Tools: Rational, Argo-UML, …
- DB-Design mit UML: Klassendiagramme
  - Modellierung in UML mit späterer Übersetzung in relationale Schemata
  - Beispiele: SP, EMBL, GIMS, ArrayExpress, …
  - Direkte Umsetzung in ORDBMS möglich

- Beispiel GIMS (Genome Information Management System): Verwendung des OODBMS „FastObjects“ (Versant)



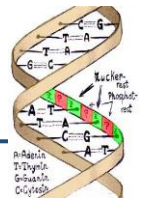
UML-Beispiel: GIMS



N.W.Paton: Conceptual modelling of genomic information, Bioinformatics, 2000.

UML: Zusammenfassung

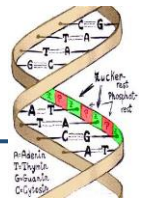
- Vorteile
  - Reichhaltiges Datenmodell mit klar definierter graphischer Notation (durch Metamodell)
  - Industriestandard für Modellierung / Entwicklung
  - Enge Verkopplung mit Software möglich (Automatische Erzeugung von Persistenzschicht: Schema plus Klassen)
- Nachteile
  - Dualität OO - RDBMS nicht trivial (Impedance Mismatch)
  - OO-Übersetzung erzeugt wenig intuitive Schema
  - Keine Anfragesprache definiert
  - Keine Untersützung von semistrukturierten Daten



XML-basierte Modelle

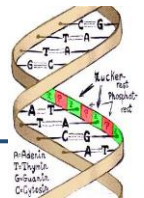
- XML: Extensible Markup Language
- Standard zur Definition von Austauschformaten
- Version 1: 1998
- SGML-basiert
- W3C-Standard
- Kern einer Sprachgruppe: XSL, Xpath, XQuery, XLink, …

```
<Feature-tables>
  <Feature-table>
    <Reference>
      <RefAuthors>
        Moore W.S., DeFilippis V.R.
      </RefAuthors>
      <RefTitle>
        The window of taxonomic resolution
      </RefTitle>
      <RefJournal>
        ...
```



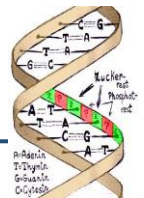
XML, DTD, XML Schema

- Document Type Definition (DTD)
  - Definition erlaubter Elemente und ihrer Attribute
  - Assoziationen durch ID, IDREF
  - Unzureichende Datentypisierung
- XML-Schema: Erweiterung
  - Constraints und Kardinalitäten
  - Vordefinierte und benutzerdefinierte Datentypen
  - Einfache und komplexe Datentypen
- XML-Dokument ist
  - Wohlgeformt: Entspricht XML Syntax
  - Gültig/valide bzgl. einer DTD: Entspricht einer gegebenen DTD
- Speicherung
  - Flatfile, XML-Datenbank (eXist, Tamino, …),
    Relationales Mapping (XML-enabled)



XML in der Bioinformatik

- GAME: Genome Annotation Markup Elements
- BIOML: BIOpolymer Markup Language
- Und viele andere DTDs und XML-Schemata für Bio-Daten
  - BSML: Bioinformatic Sequence Markup Language
  - Drosophila Genome Project
  - VisualGenomics
  - CML, OMF, DAS, CSHL, BSA, OMG-LSR
  - …



GAME

- DTD + Tools für Austausch von Genom-Annotationen
- Informationen über Sequenzabschnitte
- Ergänzt GFF-Format
  - GFF: standardisiert Darstellung der Genstrukturen
  - GAME erweitert GFF mit Metainformationen (Annotationen)

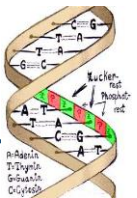
GAME Semantics

Annotation
- *"A collection of features found on an associated set of sequences"*

Features
- *"Conclusions describing intervals on different sequences. Supported by analytical evidence"*

Analyses
- *"Computer or biological experiments on a sequence. Results apply to sequence interval"*

Sequences
- *"Biological sequences in which we're interested"*

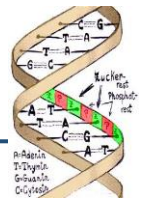


GAME DTD (Ausschnitt)

```
<!ELEMENT game ANY>
…
<!ELEMENT offset (#PCDATA)>
<!ELEMENT length (#PCDATA)>

<!ENTITY % site_operator
  " site_operator (less_than | greater_than)">

<!ELEMENT fuzzy_start (span)>
  <!ATTLIST fuzzy_start
             %site_operator; #IMPLIED>

<!ELEMENT fuzzy_end (span)>
  <!ATTLIST fuzzy_end
             %site_operator; #IMPLIED>

<!ELEMENT fuzzy_span (fuzzy_start, fuzzy_end)>

<!ELEMENT span (offset, length)>
  <!ATTLIST span
             between    (TRUE) #IMPLIED
             either_dir (TRUE) #IMPLIED>
```

```
Location: 340..565
        <span>
           <offset>339</offset>
           <length>225</length>
        </span>
```

```
Location: <345..500
        <fuzzy_span>
           <fuzzy_start site_operator="less_than">
              <span>
                 <offset>344</offset>
                 <length>1</length>
              </span>
           </fuzzy_start>
           <fuzzy_end>
              <span>
                 <offset>499</offset>
                 <length>1</length>
              </span>
           </fuzzy_end>
        </fuzzy_span>
```



GAME: Pfam Beispiel

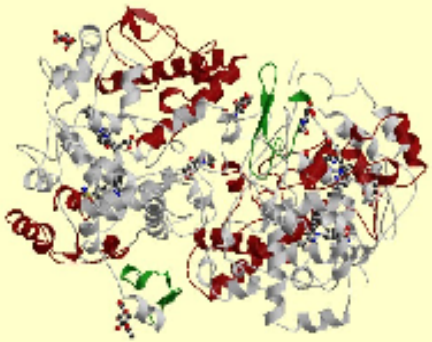


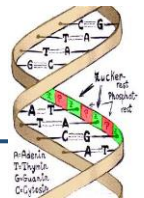
Figure 1: 6cox
Oxidoreductase
Cyclooxygenase-2 (prostaglandin synthase-2) complexed with a selective inhibitor, sc-558 in i222 space group

Key:

Domain	Chain	Start Residue	End Residue
An_peroxidase	A	228	344
EGF	A	36	69
An_peroxidase	A	464	521
An_peroxidase	B	228	344
EGF	B	36	69
An_peroxidase	B	464	521

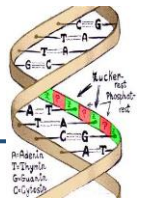
Beispiel für Protein-Domain: EGF-like domain
(Pfam-ID PF00008)

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown PUB00001077, PUB00001077, [MEDLINE:84117505], [MEDLINE:91145344], [MEDLINE:85063790], PUB00004964 to be present, in a more or less conserved form, in a large number of other, mostly animal proteins.
The list of proteins currently known to contain one or more copies of an EGF-like pattern is large and varied. The functional significance of EGF domains in what appear to be unrelated proteins is not yet clear. However, a common feature is that these repeats are found in the extracellular domain of membrane-bound proteins or in proteins known to be secreted (exception: prostaglandin G/H synthase). The EGF domain includes six cysteine residues which have been shown (in EGF) to be involved in disulphide bonds. The main structure is a two-stranded -sheet followed by a loop to a C-terminal short two-stranded sheet. Subdomains between the conserved cysteines vary in length.



GAME: XML-Darstellung von Pf00008

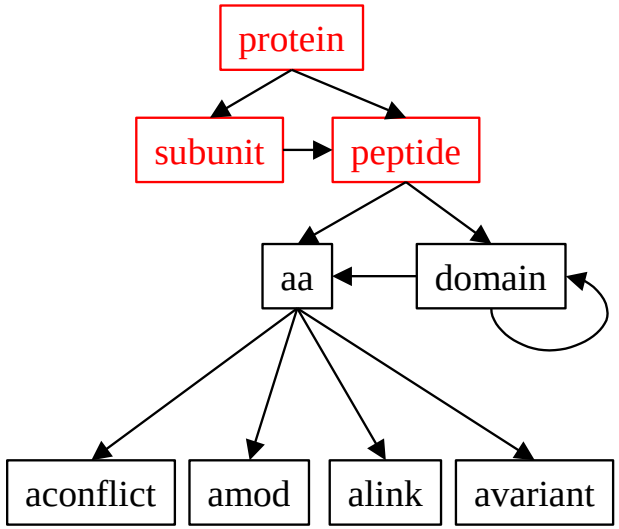
```
<computational_analysis seq="dmNotch">
<date>08/26/1999</date> <program>hmmpfam</program> <version>2.1.1</version>
<database>
          <name>Pfam</name> // Bezug auf "Pfam" (Protein families database of
                            // alignments and HMMs
          <date>god (and Sean Eddy) knows when it was created</date>
          <version>4.1</version>
</database>
<result_set>
          <dbxref>
          <database> <name>Pfam</name> </database>
          <unique_id>PF00008</unique_id> // Pfam accession number
          </dbxref>
          <output> <type>Description</type><value>EGF-like domain</value> </output>
          …
          <result_span>
             <score> 22.6 </score>
             <type>Motif</type>
             <subtype>EGF</subtype> // EGF: epidermal growth factor (family)
             <seq_relationship seq="dmNotch" type="query">
                    <span>
                       <offset>62</offset>
                       <length>32</length>
                    </span>
                    <alignment>CTSV-GCQNGGTCVTQLN------GKTYCACDSH-----YVGDYC</alignment>
             </seq_relationship>
             <seq_relationship seq="EGF" type="subject">
                    <span>
                       <offset>0</offset>
                       <length>44</length>
                    </span>
                    <alignment>CapnnpCsngGtCvntpggssdnfggytCeCppGdyylsytGkrC</alignment>
             </seq_relationship>
           </result_span>
</result_set>
</computational_analysis>
```

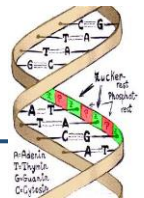


BIOML

- BIOpolymer Markup Language;
  entwickelt von ProteoMetrics (→ Genomic Solutions)
- Austauschformat für Experimentresultate bzgl. Biopolymeren (61 XML-Eelemente)

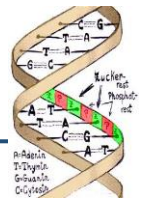
```
<!ELEMENT protein (#PCDATA|subunit|peptide|%nr;)*>
  <!ATTLIST protein %global; %comp;>

<!ELEMENT subunit (#PCDATA|peptide|%nr;)*>
  <!ATTLIST subunit %global; %comp;>

<!ELEMENT peptide (#PCDATA|domain|aa|%nr;)*>
  <!ATTLIST peptide  %global; %start; %end;>

<!ELEMENT domain (#PCDATA|domain|aa|%nr;)*>
  <!ATTLIST domain %global; %start; %end; %dom_type;>

<!ELEMENT aa (#PCDATA|amod|alink|avariant|aconflict|%nr;)*>
  <!ATTLIST aa %aa_type; %global; %at; %to;>

<!ELEMENT amod (#PCDATA|%nr;)*>
  <!ATTLIST amod %type; %global; %at; %to; %occupied; %comp; %covalent;>

<!ELEMENT alink (#PCDATA|%nr;)*>
  <!ATTLIST alink %type; %global; %at; %to; %occupied;>

<!ELEMENT avariant (#PCDATA|%nr;)*>
  <!ATTLIST avariant %global; %aa_type; %at; %occupied;>

<!ELEMENT aconflict (#PCDATA|%nr;)*>
  <!ATTLIST aconflict %global; %aa_type; %at;>
```





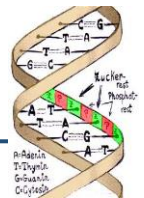
XML in der Bioinformatik: Bewertung

- Gut geeignet für semi-strukturierte Biodaten
- Vorteile (insb. gegenüber Entry-basiertem Modell)
  - Industriestandard, viele Tools (Editoren)
  - DTD generierbar aus UML-, Java-Spezifikationen, …
  - Effiziente Parser
  - Unterstützung durch relationale DB-Hersteller (IBM, Oracle, …)
  - Zunehmend XML-Datenbanken verfügbar (eXist, Tamino etc.)
  - Strukturierte Anfragen (XQuery) und Textsuche möglich
- Nachteile
  - Dokumente sehr lang, daher oft nicht sehr gut lesbar
  - Ohne DTD: Keine Dokumentvalidität, keine Semantik für Datenaustausch
  - Mit DTD: Geringere Flexibilität, Dokumente evtl. ungültig bei Änderungen



Zusammenfassung

- "Austauschformate"
  - Entry-based
  - ASN.1
  - XML
- Speichern und Anfragen
  - Relationales Modell
  - Objektorientiertes/Objektrelationales Modell
- Vorteile der Flatfiles nicht unterschätzen
  - Viele Bio-Einrichtungen ohne RDBMS / Informatiker
- I.d.R. mehrere Formate/Datenmodelle in *einem* Bio-Projekt



Fragen ?