

バイオインフォマティクスとデータグリッド

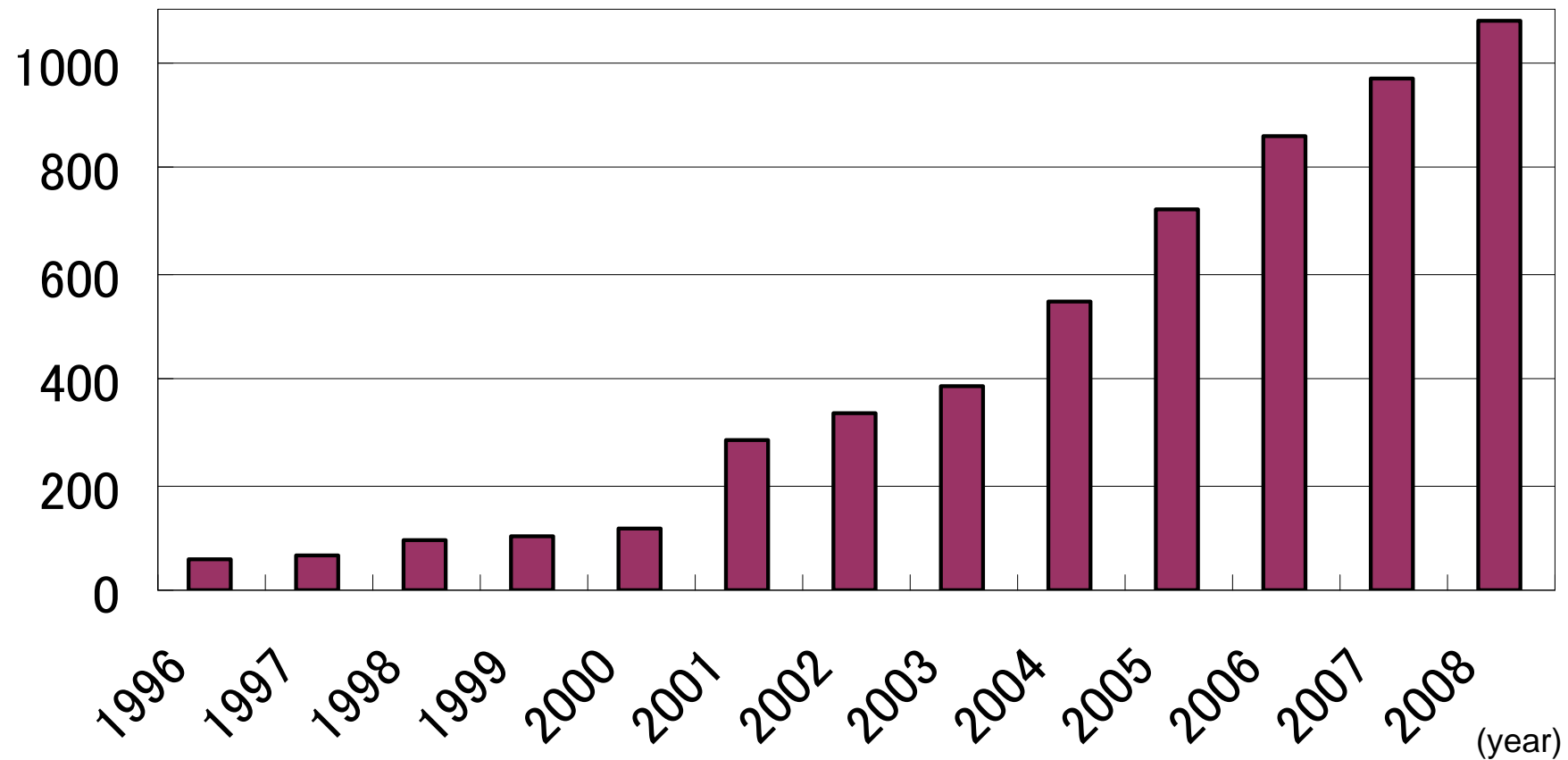
松田秀雄

(大阪大学 大学院情報科学研究科)

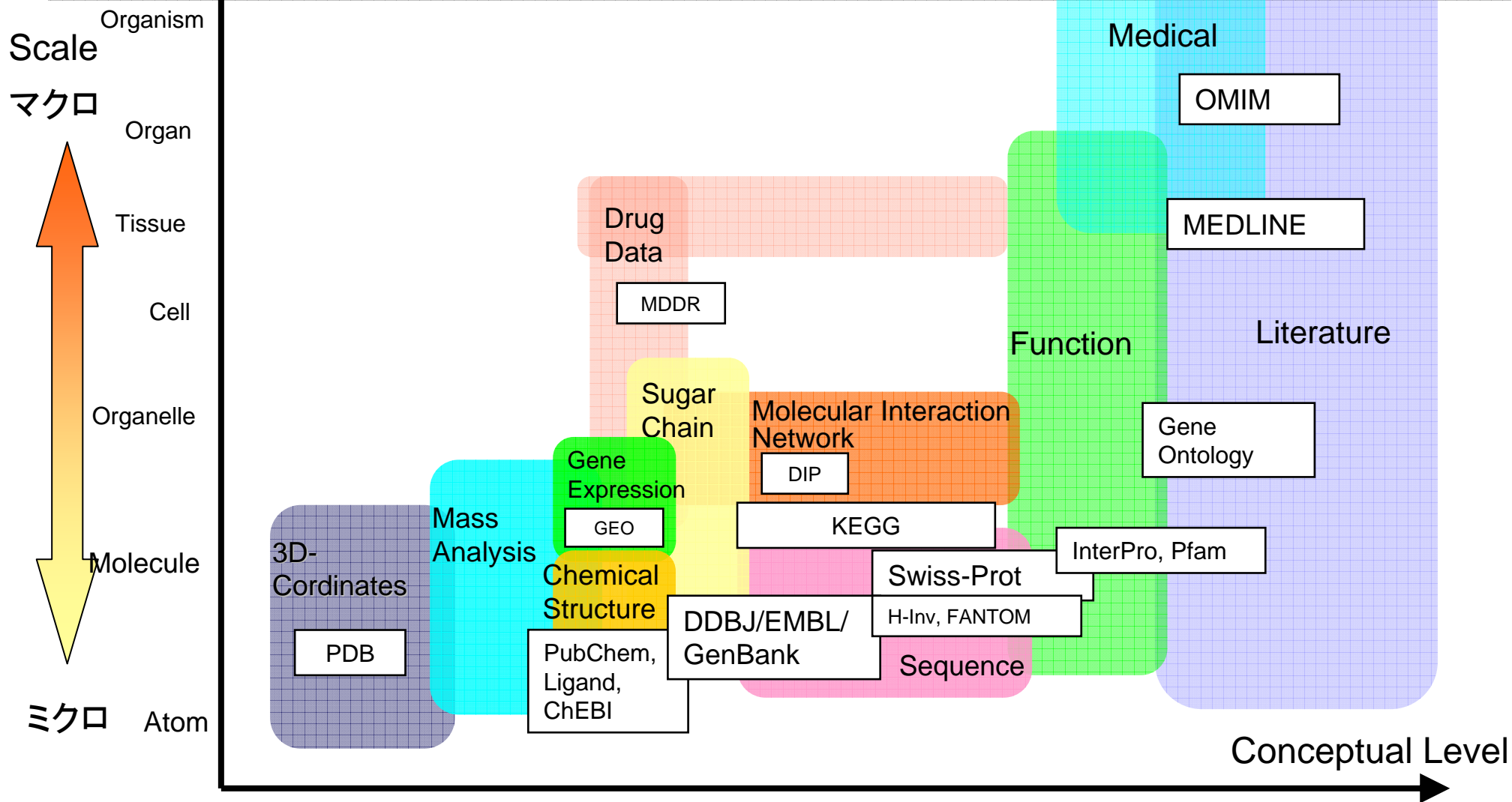
生物学の実験技術の進歩によるデータの増加

- 例えば分子生物学のデータベースでは
 - 1078個のデータベースが公開(2007年の1年間で新規に110個のデータベースが追加、Nucleic Acids Research誌DB特集から)
- なぜ、そんなにデータベースが増えていくのか？
 - 取り扱う対象や実験技術の進歩
 - 対象:ゲノム(DNA)、プロテオーム(タンパク質)、メタボローム(酵素反応)、セルローム(細胞)、フィジオーム(臓器・生理現象)、バーチャルヒューマン
 - 実験技術:DNA配列決定、リアルタイムPCR、タンパク質質量分析、タンパク質相互作用、バイオチップ、...
 - 解析ツールの増加(1000種類以上ある?)
 - FASTA / BLAST / BLAT, ClustalW / DiAlign / Meme,

- ヒトゲノム解読完了(2001年)から急激に増加



Nucleic Acids Research DB issueから抽出



詳細な記述
(数値データ)



概念的な記述
(文献など)

- 異種性が大きい原因
 - 分子種の違い
 - DNA, タンパク質, 糖および糖鎖, 膜(脂質), その他の低分子化合物など
 - コミュニティの違い
 - 微生物／植物／動物
 - 生化学／生物物理学／バイオインフォマティクス
 - 基礎生物学／臨床・医療・保健
 - 共通の表記法の欠如
 - 特定の生物種の遺伝子名(例えば、ヒトの遺伝子名はHUGOが決められている)については標準化がされているが、生物種を超えるとまちまち。
 - メタデータがデータベースごとに異なる(極論すると、研究者ごとに異なる)
 - 標準化することのメリットが認識されていない?
- メタデータの概念レベルでの標準化の活動はある
 - Gene Ontology(遺伝子機能を表現する語彙を統制する)



データベースの形 データの形 タンパク質アミノ酸配列

DNA 塩基配列(DDBJ)

(SWISS-PROT)

タンパク質立体構造(PDB)

LOCUS ECORBS 6197 bp DNA ..
30-OCT-1994

DEFINITION E. coli ... rbsB
transport system, ...

ACCESSION M13169 M13517

NID g147511

KEYWORDS high affinity ribose ..

SOURCE E. coli K12 DNA.

ORGANISM Escherichia coli

REFERENCE 1 (sites)

AUTHORS ...

TITLE ...

JOURNAL ...

MEDLINE 84032513

FEATURES

gene 122..127
/gene="rbsB"
/translation="MKKGTVLNS...
...

ORIGIN 94 bp upstream of BclI
1 ctcaggttcg aatctaac.

ID RBSB_ECOLI ...

AC P02925;

DT 21-JUL-1986 (... CREATED)

DT 21-JUL-1986 (... UPDATE)

DT 01-NOV-1995 (... UPDATE)

DE D-RIBOSE-BINDING ...

GN RBSB OR RBSP OR PRLB.

OS ESCHERICHIA COLI.

OC PROKARYOTA; GRACILICUTES;
OC ENTEROBACTERIACEAE.

RN [1]RP SEQUENCE FROM N. A., ...

RX MEDLINE; 84032513.

RA GROARKE J. M., MAHONEY W. C.,
RA ZALKIN H., HERMODSON M. A. ;
RL J. BIOL. CHEM. 258:...

CC -!- FUNCTION: INVOLVED IN
...
KW TRANSPORT; SUGAR TRANSPORT;
KW 3D-STRUCTURE.

SQ SEQUENCE 296 AA; ...
MNMKKLATLV SAVALSATVS ANA...
....

HEADER SUGAR TRANSPORT 23-SEP-94

COMPND D-RIBOSE-BINDING PROTEIN

COMPND 2 (G134R) COMPLEXED WITH

SOURCE (ESCHERICHIA COLI)...

SOURCE 2 EXPRESSION PLASMID)

AUTHOR S. L. MOWBRAY, A. J. BJORKMAN

REVDAT 1 26-JAN-95 1DRJ 0

JRNL AUTH A. J. BJORKMAN

JRNL TITL PROBING PR...

JRNL TITL 2 RIBOSE-BINDING

JRNL REF TO BE PUBLISHED

JRNL REFN ASTM

SEQRES 1 271 LYS ASP THR ...

SEQRES 2 271 PRO PHE PHE ...

...

HELIX 1 A PRO 14 LEU ...

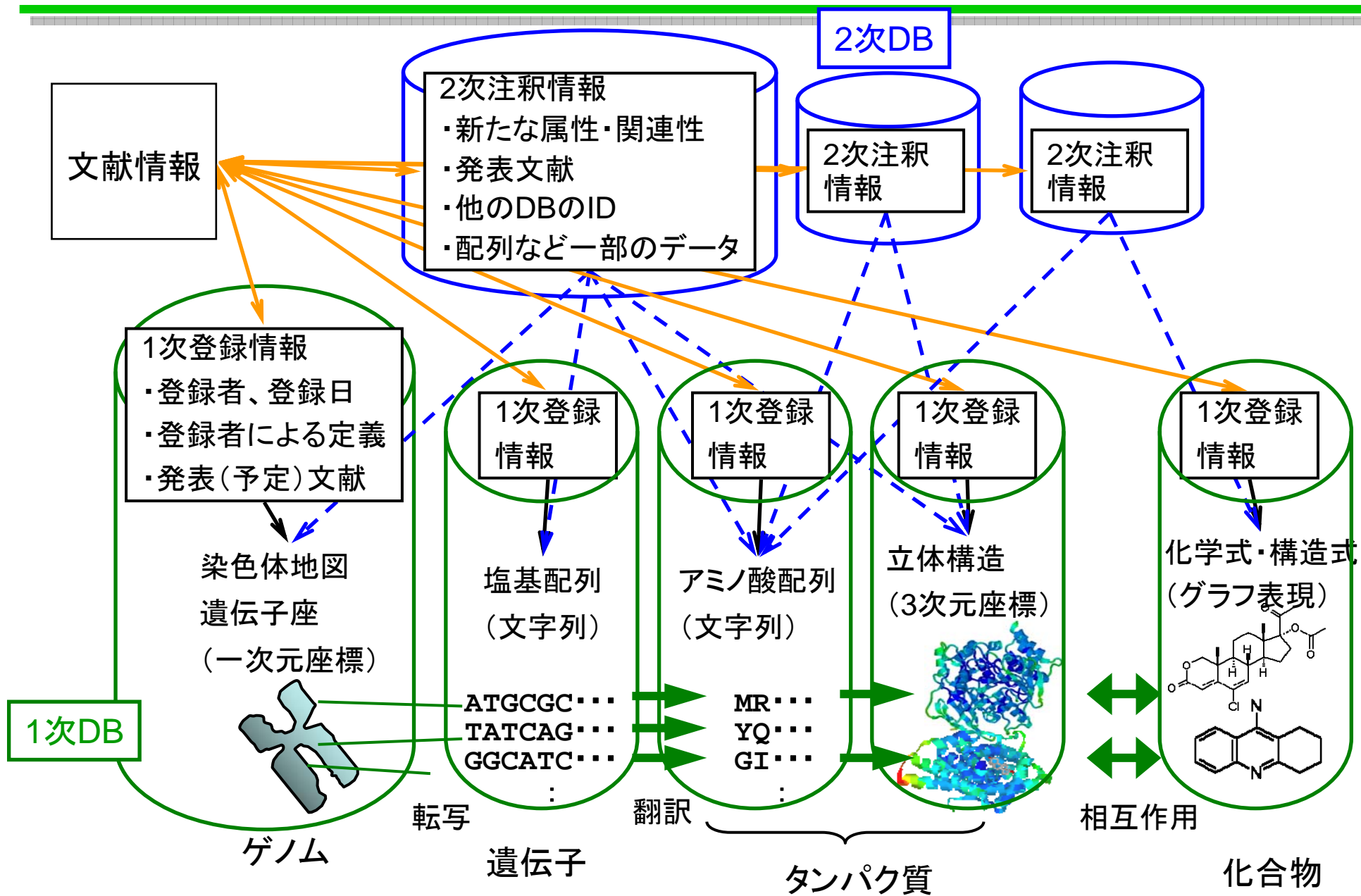
HELIX 2 B PRO 43 LEU ...

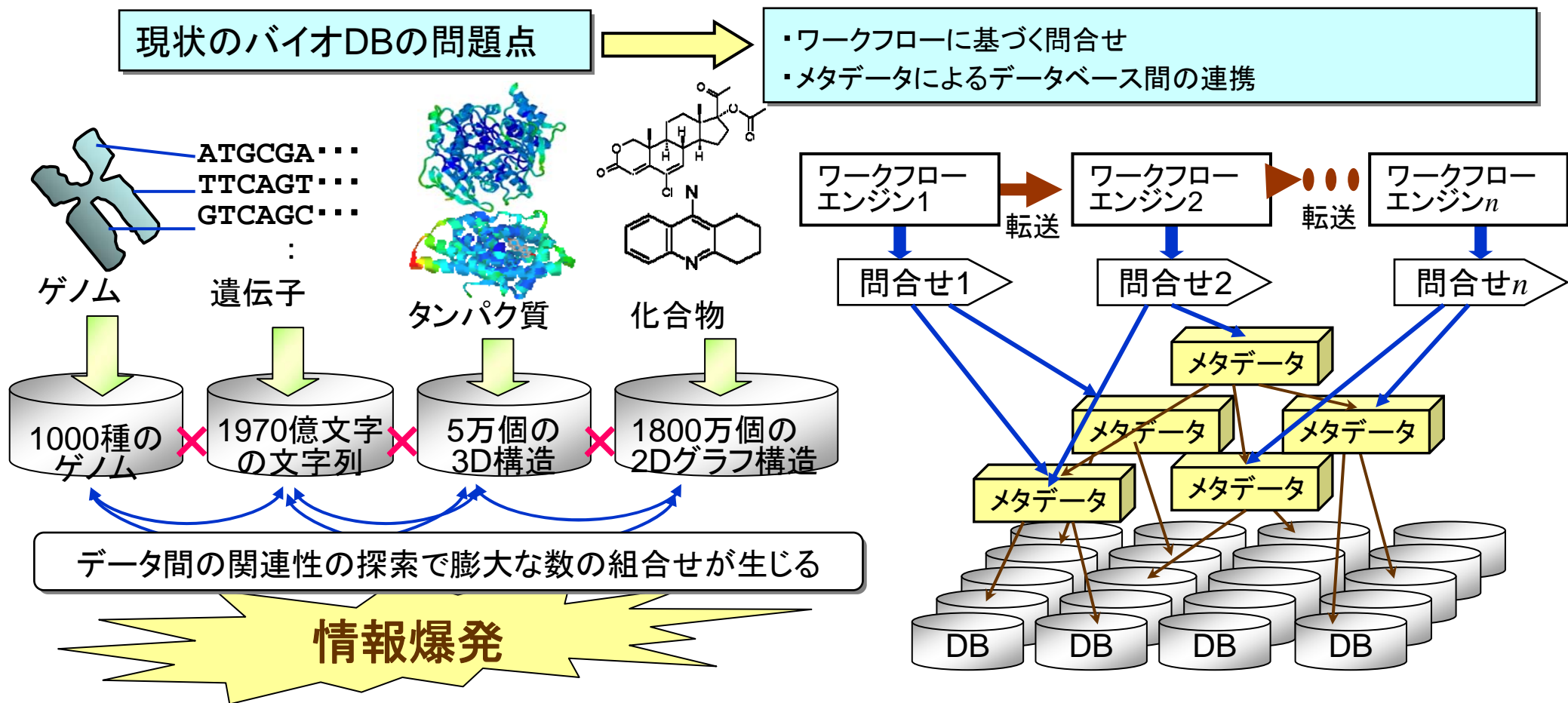
...

ATOM 1 N LYS 1 x1 y1 z1

ATOM 2 CA LYS 1 x2 y2 z2

...





- 統合データベースプロジェクトについて

(総合科学技術会議資料より)

- 統合の対象

- ライフサイエンス関連の分子データ、文献データ、臨床などの表現型データ

- 提供する機能(抜粋)

- データベースやツールの所在や利用法を網羅したポータルサイト
- 分子データと文献知識(高次生命機能)を統合したデータベース
- 分野の俯瞰や仮説生成が容易に行える検索機能

- 委託実施機関

- 中核機関 情報・システム研究機構 ライフサイエンス統合データベースセンター
- 分担機関 京大、東京医科歯科大学、東大

- 本研究

- 異種データベース、データグリッドなどの技術に基づき、広域に分散したバイオデータベース群からの情報統合を通じて、データ統合の基盤技術を開発

- 最もデータ量の大きいのは、DNA塩基配列
- DDBJ(日本)、EMBL(EU)、GenBank(US)の3箇所で相互にデータを交換しながらデータベースを管理
- 最新のデータ量の統計情報
 - 塩基配列数 197G bp、エントリ数 112M
- 更新には3種類がある
 1. 差分データの更新: 前回のリリースからの更新であり、毎日更新される
単純に挿入されるだけでそれまでのデータは変更なし
データ量(フラットファイルと呼ばれるテキストファイル形式をgzipで圧縮した形式)平均76MB(最小2MB、最大736MB)
 2. リリースの更新(年4回): 差分データの取り込み以外に、既存データの更新を含む
 3. データベースの形式の更新(年2回): 基本的には、細かいデータ項目(実質的にメタデータ部分)の更新だが、中には大きな変更もある

- 生データ(例えばDNA塩基配列)だけを見ても、ほとんど情報がないので、データに対する注釈情報としてのメタデータが必須
 - 生物種、分類カテゴリ
 - DNAかRNAかの区別
 - DNA塩基配列上での遺伝子の開始位置、終了位置
 - タンパク質をコードしている遺伝子の場合は、翻訳したアミノ酸配列
 - 発表文献
 - キーワード
- メタデータは、データベースの管理機関が独自に決めている
- メタデータは、データベースの構造中に密に組み込まれており、生データとメタデータの分離は困難

- 現状のメタデータは、データベースの構造やデータ形式に密に組み込まれており、大きな変更は困難
- 現状のデータ形式を、e-Scienceからの提案で変更することはほぼ不可能
- しかし、データベースを横断的に検索するには、個々のデータベースの間でのデータの対応付けが必要不可欠(特定のデータベースの構造に依存しないメタデータが必要)
- メタデータの表現にはXMLが有効
 - 名前空間を複数持てる
 - 関係ないタグは無視することが可能



XMLメタデータの試作

```
<?xml version="1.0"?>
<entry>
  <id>データベースID</id>
  <date>登録の日付</date>
  <name db="pir" acc=".." >タンパク質名</name>
  <gene db="sp" acc="..">遺伝子名</gene>
  <organism db="pir" acc="..">
    <scientific db="pir" acc="..">生物種の学名</scientific>
    <common db="pir" acc="..">生物種の慣用名</common>
  </organism>
  <reference db="pir" acc="..">文献情報
    <author>著者</author> <citation>学術誌名</citation> <volume>巻</volume>
    <year>発表年</year> <title>文献題名</title>
    <first_page>開始ページ</first_page> <last_page>終了ページ</last_page>
  </reference>
  <function db="sp" acc=".."> <desc>タンパク質の機能記述</desc> </function>
  <keyword db="sp" acc="..">キーワードリスト</keyword>
  <feature db="pir" acc=".."> <type>部分構造のタイプ(活性部位など)</type> <desc>部分構造の
    説明</desc>
    <start>開始位置</start> <end>終了位置</end> </feature>
  <sequence db="sp" acc=".." length=".."> タンパク質の配列データ </sequence>
</entry>
```



XMLメタデータの表示例

http://dg1.ics.es.osaka-u.ac.jp/cgi-bin/search.cgi?gene...

ファイル(E) 編集(E) 表示(V) お気に入り(A) ツール(T) >> リンク >>

ENTRY1

SP	id	HGF_HUMAN
PIR	id	JH0579
PDB	id -1	2HGF
PDB	id -2	1BHT
SP	accession	P14210
		Q9B YL9
		Q9UDU6
PIR	accession	JH0579
		JU0333
		A41140
		B36677
		A36677
		A33512
		A39006
		PH0114
		A37796
		S06794
		I59214
S15443		
I52253		
PDB	accession -1	2HGF
PDB	accession -2	1BHT

3種類のデータベースの検索結果を、タンパク質ごとにまとめて表示



同時に複数のデータベースの内容を比較可能

http://dg1.ics.es.osaka-u.ac.jp/cgi-bin/search.cgi?gene=HGF&Format=html ...

ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H) リンク

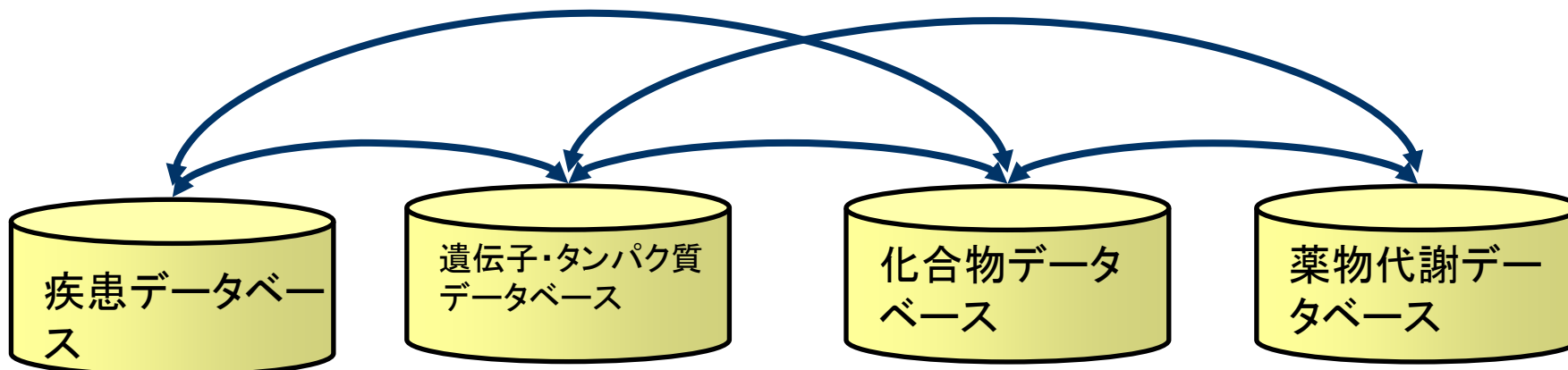
アドレス(D) http://dg1.ics.es.osaka-u.ac.jp/cgi-bin/search.cgi?gene=HGF&Format=html 移動

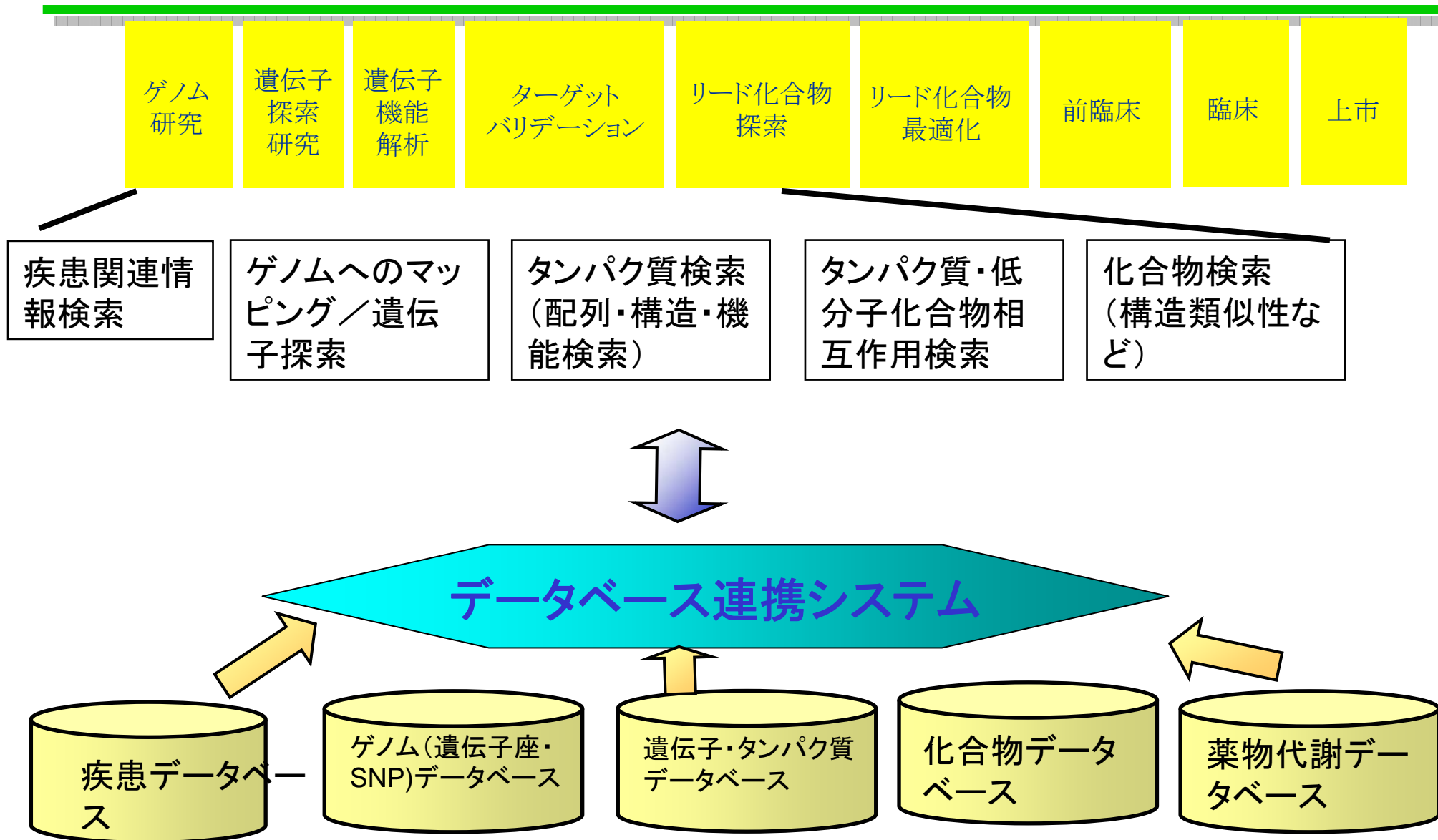
	primary	secondary
SP	keyword	TRYPSIN_DOM
		Growth factor
		Kringle
		Glycoprotein
		Serine protease homolog
		Repeat
		Signal
		3D-structure
		Polymorphism
		Pyrrolidone carboxylic acid
PIR	keyword	alternative splicing
		glycoprotein
		growth factor
		heterodimer
		kringle
		pyroglutamic acid
PDB	keyword -1	HEPATOCTE GROWTH FACTOR
		SCATTER FACTOR
		HAIRPIN LOOP
		HEPARIN BINDING
		PLASMINOGEN RELATED
		NK1

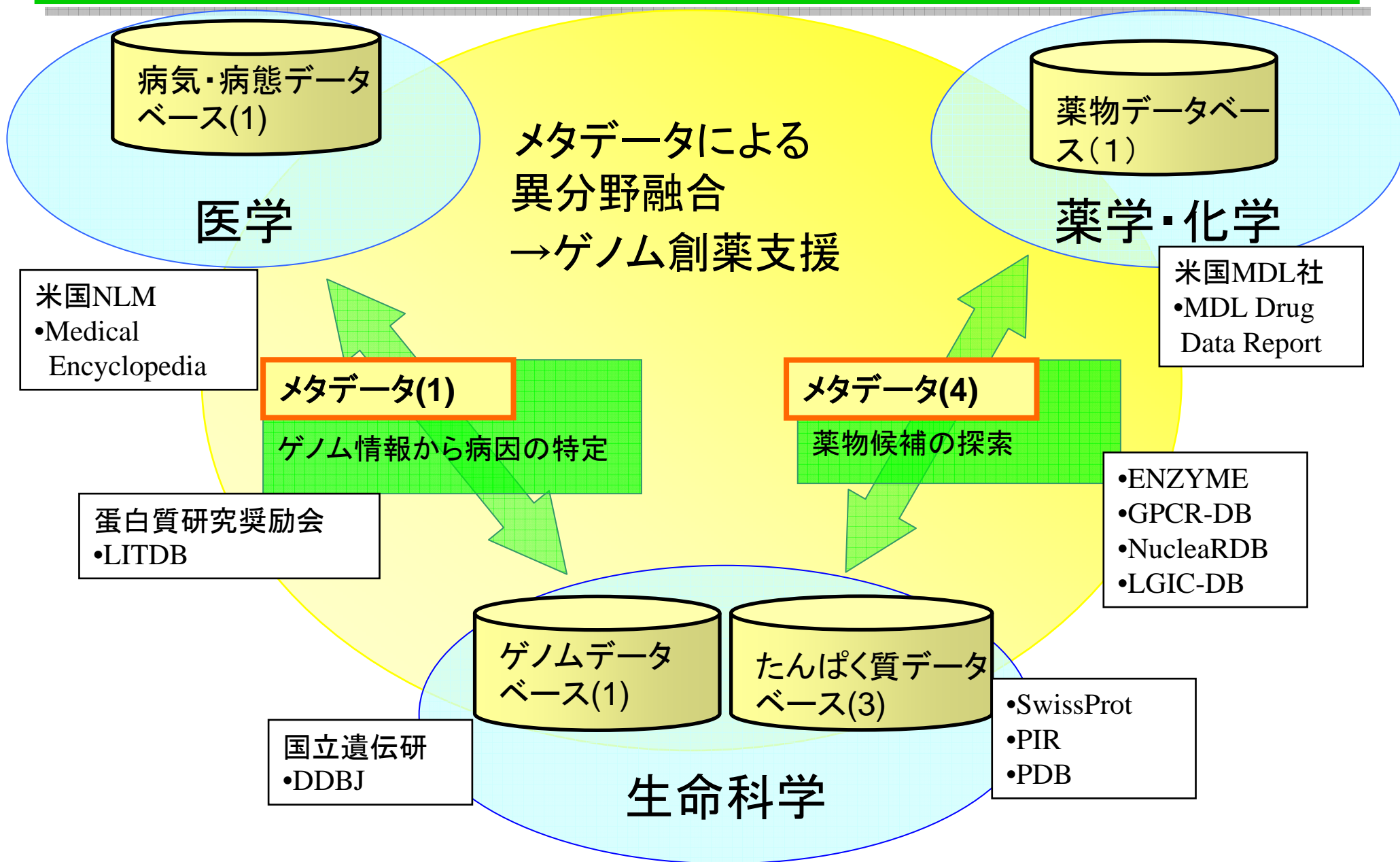


アプリケーションへの負担

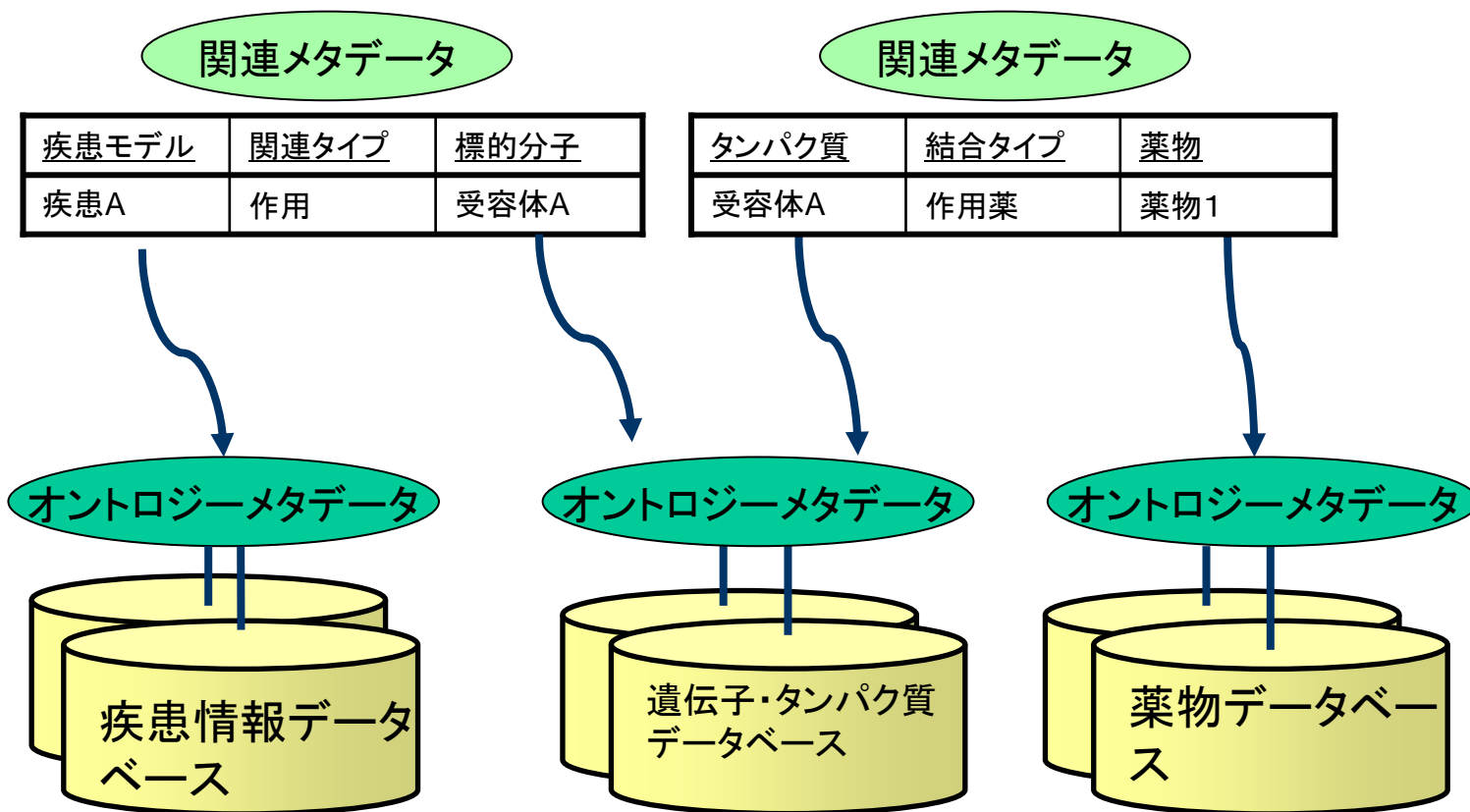
多様な関連性→膨大な組合せ
用語の違い、表記のゆれ

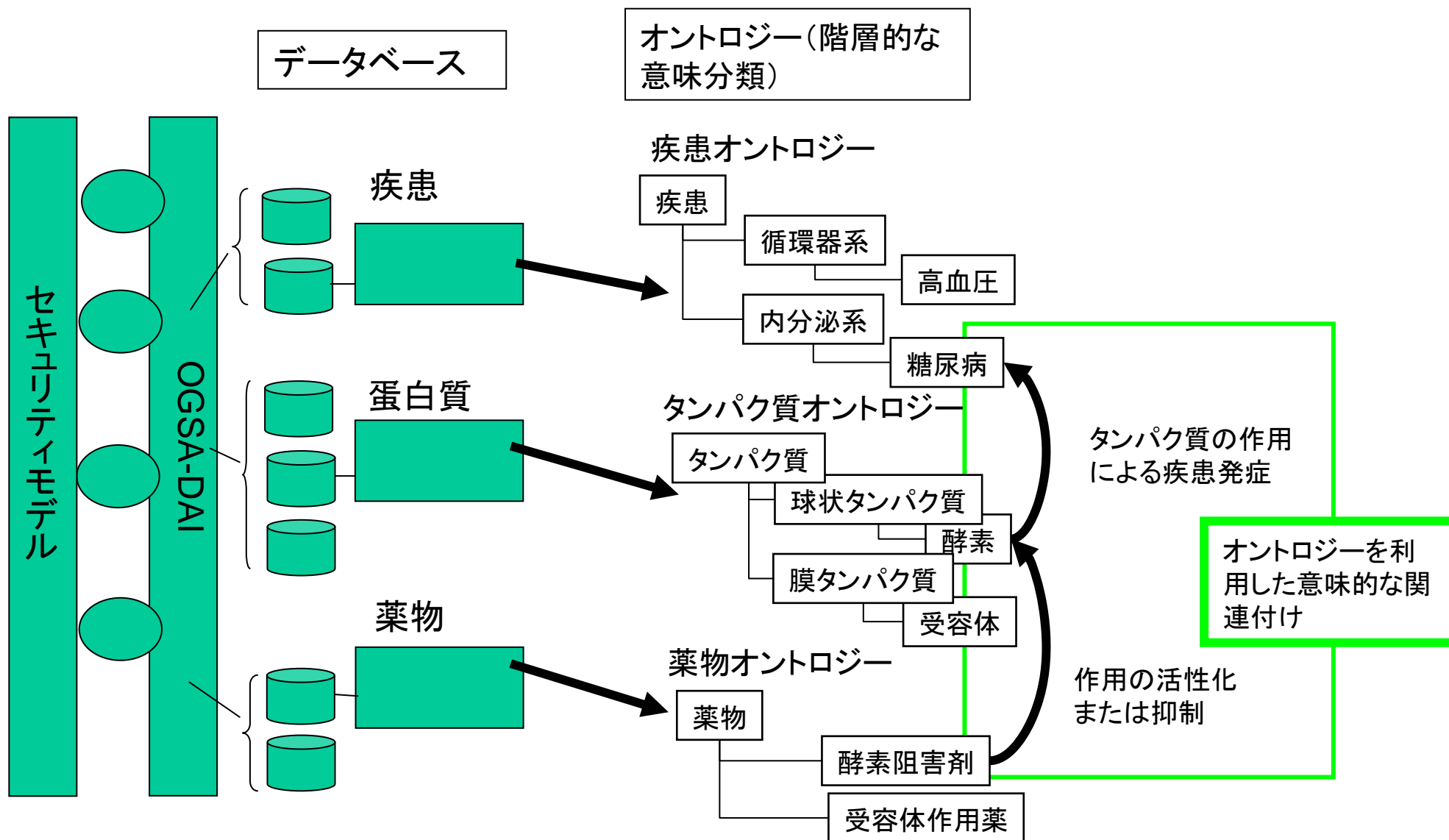




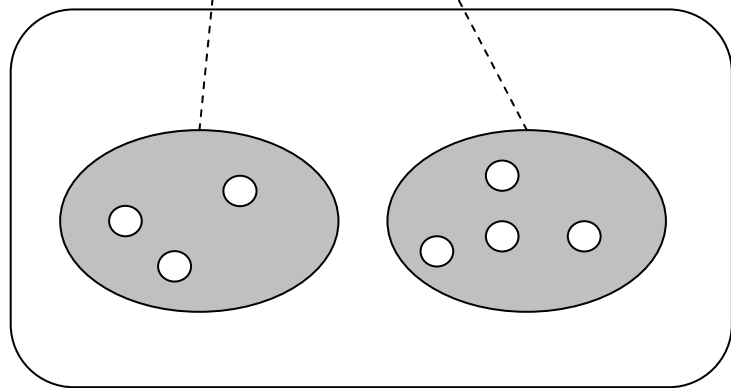
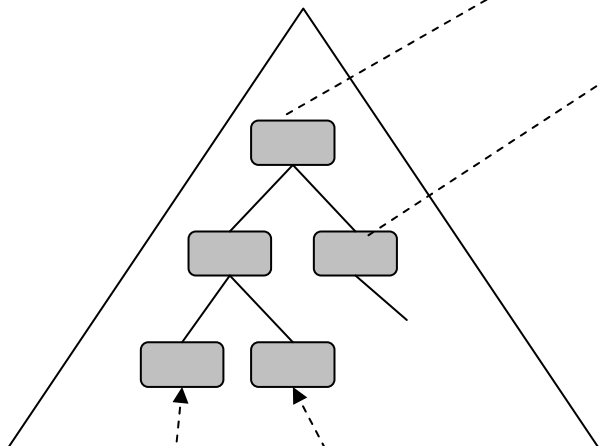


- 2階層のメタデータを介してデータベースを参照する
 - オントロジーメタデータ: オントロジーにより概念体系を明確化し、各分野の実体(例: タンパク質、薬物など)ごとの用語の違いやあいまい性を吸収
 - 関連メタデータ: 実体間の意味的な関連を記述
- メタデータ導入の利点
 - 複数のデータベースで、データ間の関連全てを直接記述すると、データ数の組合せに比例する膨大な数になる(組合せ爆発)
 - メタデータを介することにより、データ間の対応を階層的な分類間の関連にまとめることができる





タンパク質機能
オントロジー

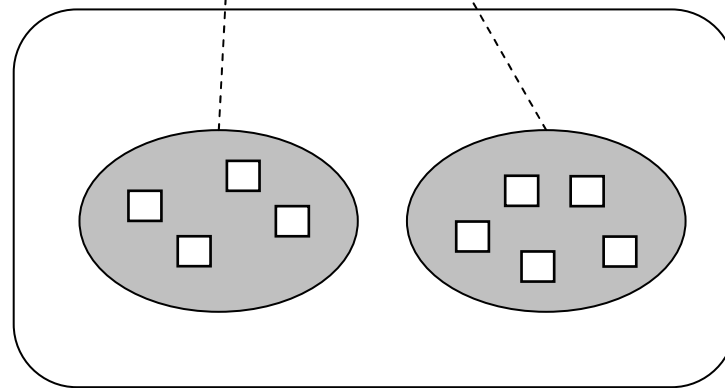
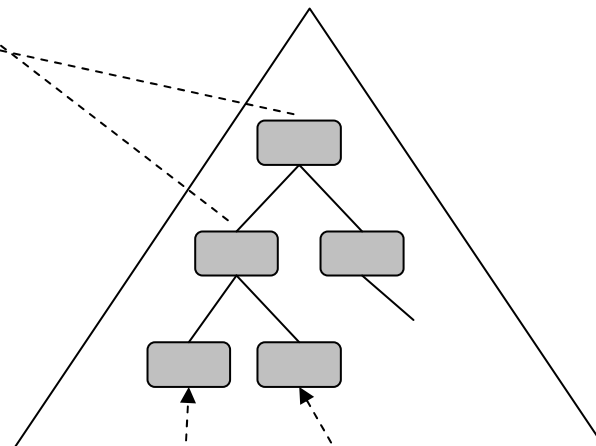


タンパク質

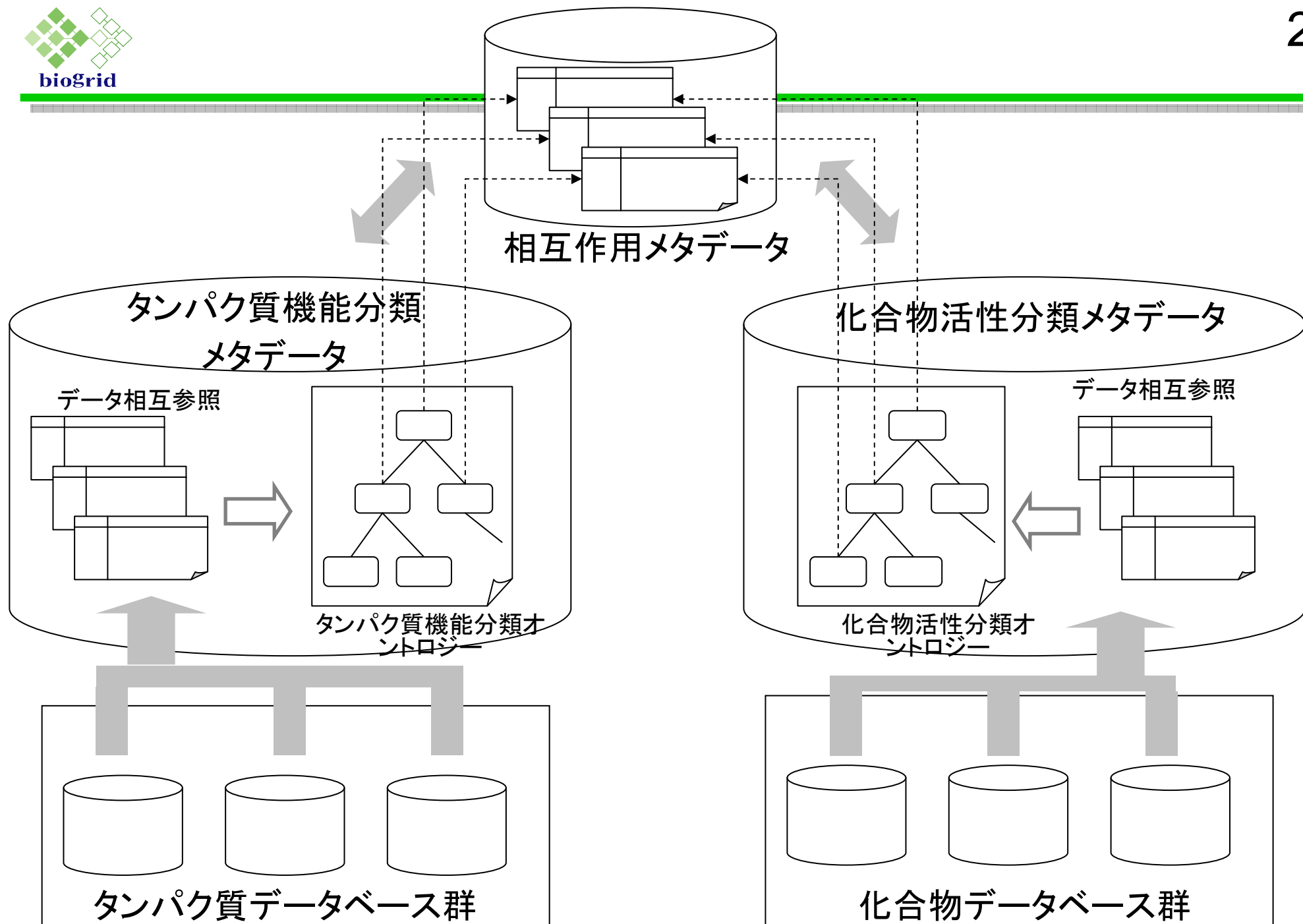
作用薬
拮抗薬
阻害薬

タンパク質・化合物
相互作用

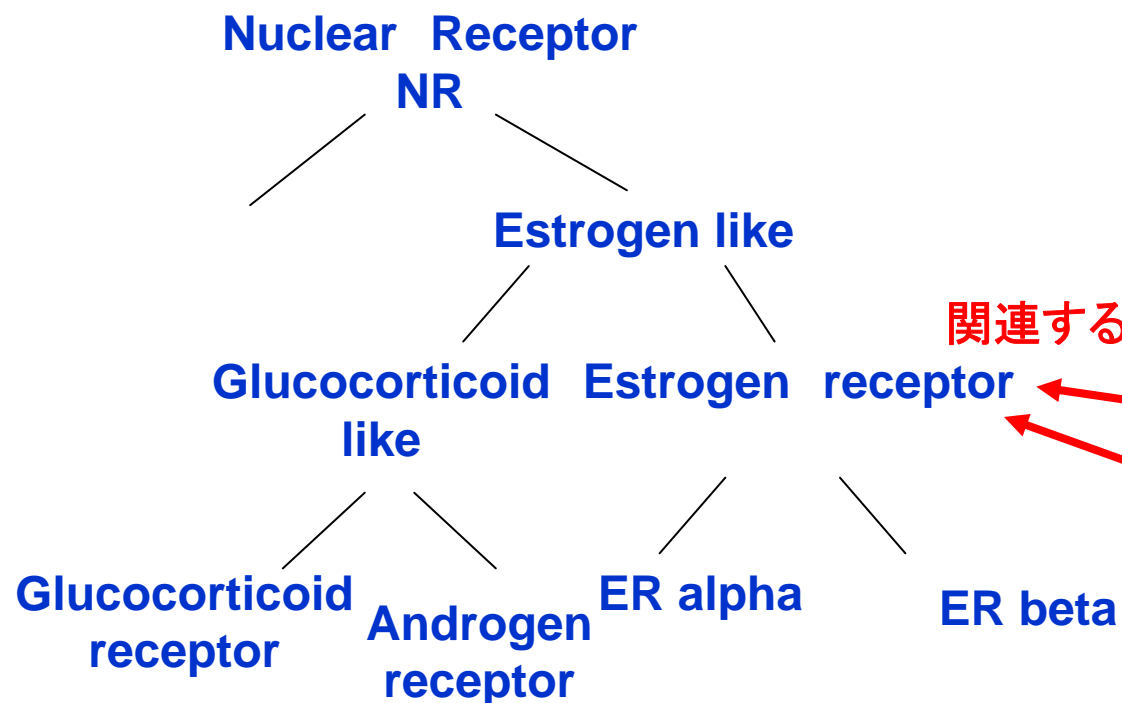
化合物活性
オントロジー



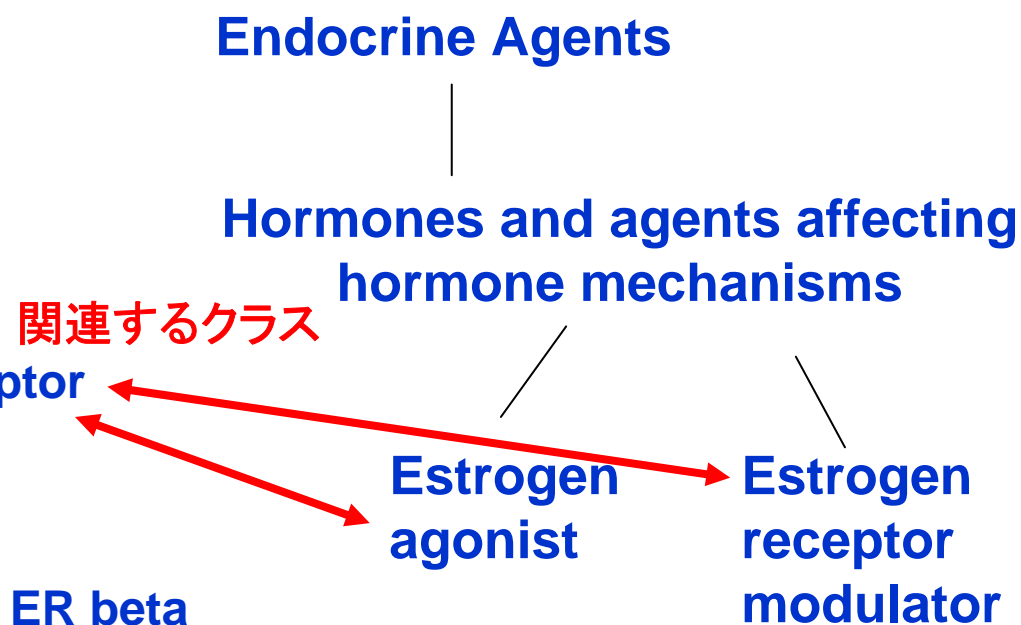
化合物



タンパク質の機能分類 (Gene Ontology)



化合物の活性分類 (MDL Drug Data Reportの活性情報から抽出)



- タンパク質・化合物を分類したオントロジーで、対応可能な概念クラスを階層構造をたどりながら探索し、クラス間の関連を抽出する
- 実際の分子間相互作用データで検証していく

疾患

Medical Encyclopedia

3K entries

PRF DB

(蛋白質研究奨励会)

パスウェイ

(生体分子間
ネットワーク)

KEGG

18K entries

タンパク質

SwissProt 163K entries,
60M amino acids

PIR 283K entries,
96M amino acids

PDB 28K protein structures

ゲノム

Ensembl, DDBJ

Human 24K genes

3.3G nucleotides

化合物

MDDR 150K compounds

PubChem, Ligand Info

1.8M compounds

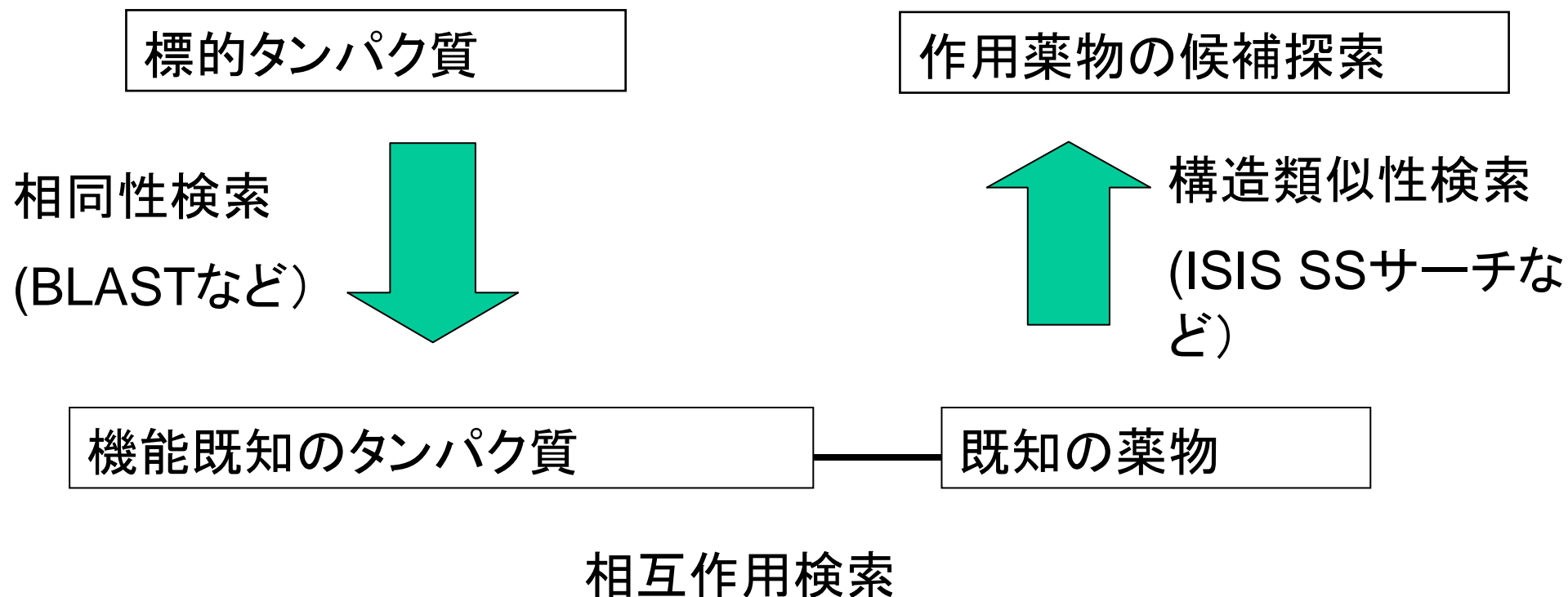
相互作用

ENZYME

GPCR-DB

NucleaRDB

LGIC DB



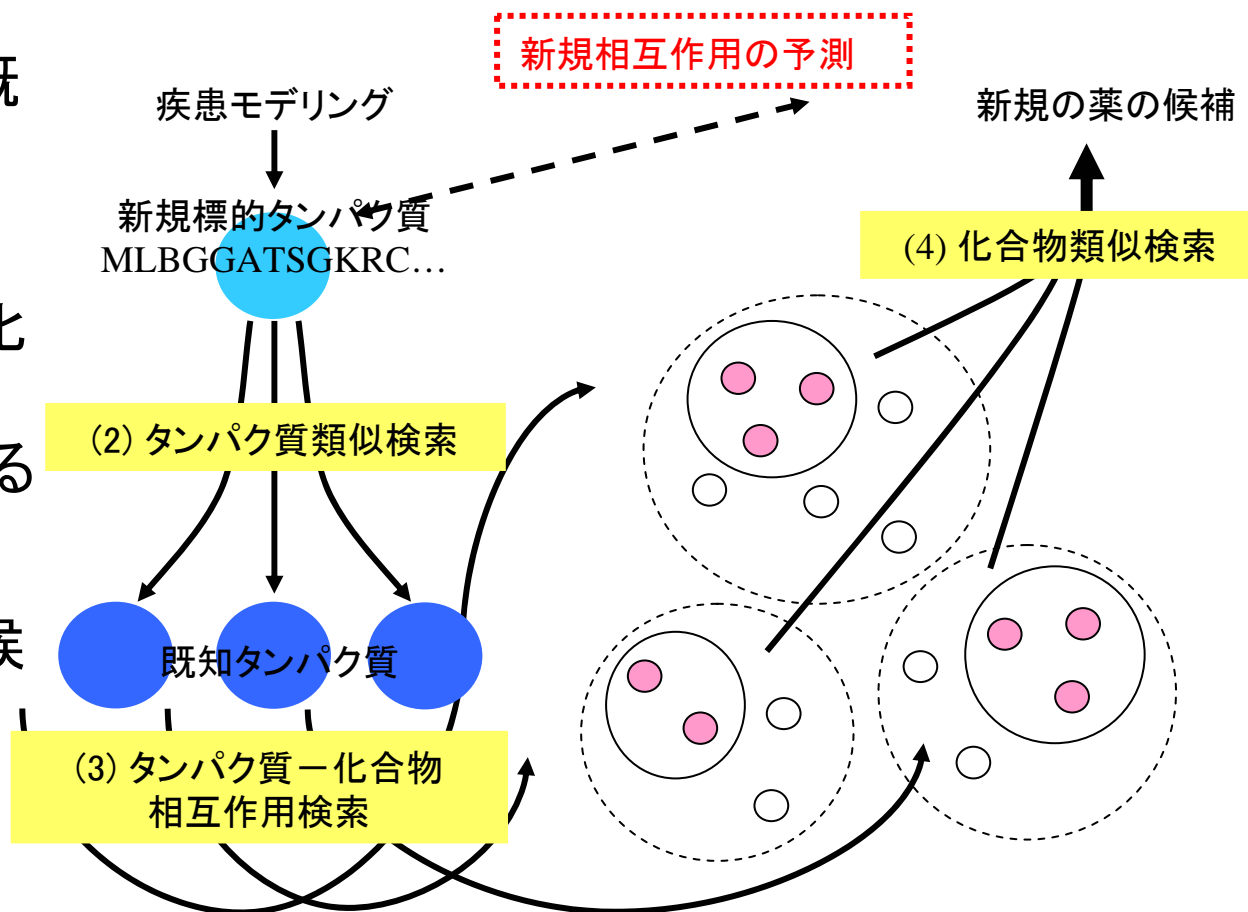
Schuffenhauer A, Floersheim P, Acklin P, Jacoby E.,
“Similarity metrics for ligands reflecting the similarity of the target proteins”,
J Chem Inf Comput Sci. 2003 Mar-Apr;43(2):391-405.

1. 疾患に関する新規の標的タンパク質を検索する

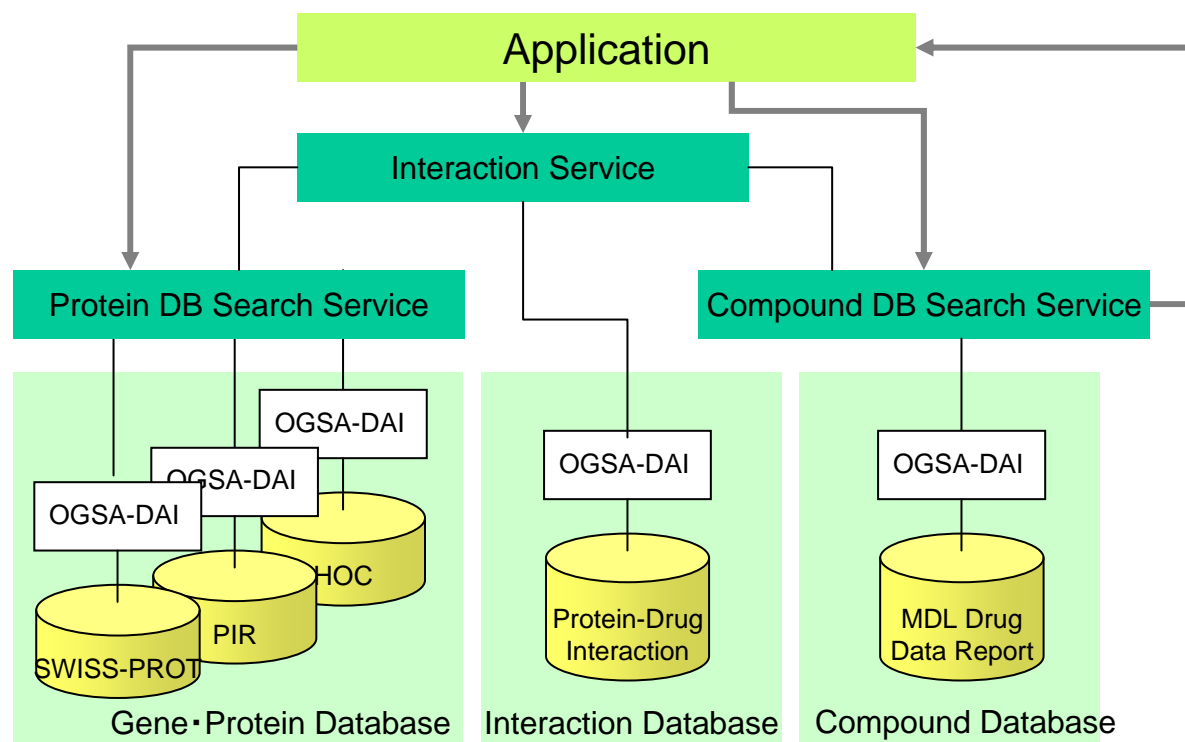
2. 標的タンパク質に類似した既知タンパク質を、BLASTなどを用いて検索する

3. 既知タンパク質に結合する化合物を、タンパク質と化合物の相互作用データから求める

4. 化合物類似検索を用いて検索された化合物から、薬の候補となる化合物を検索する

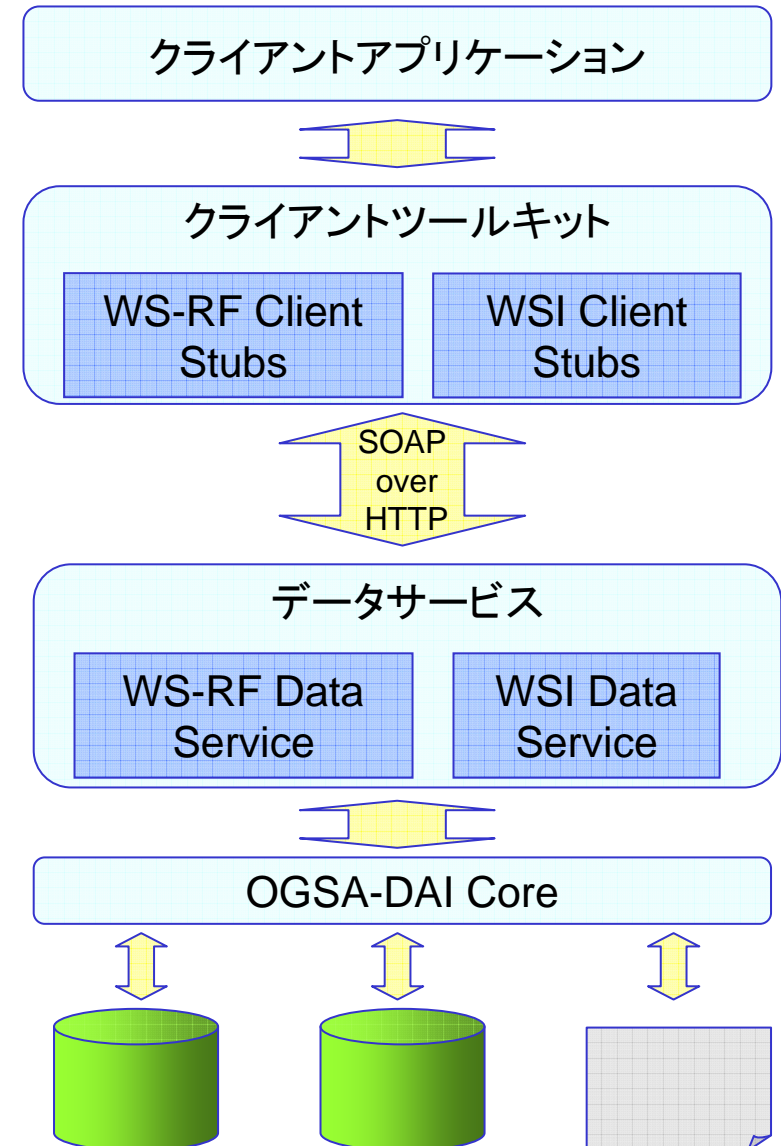


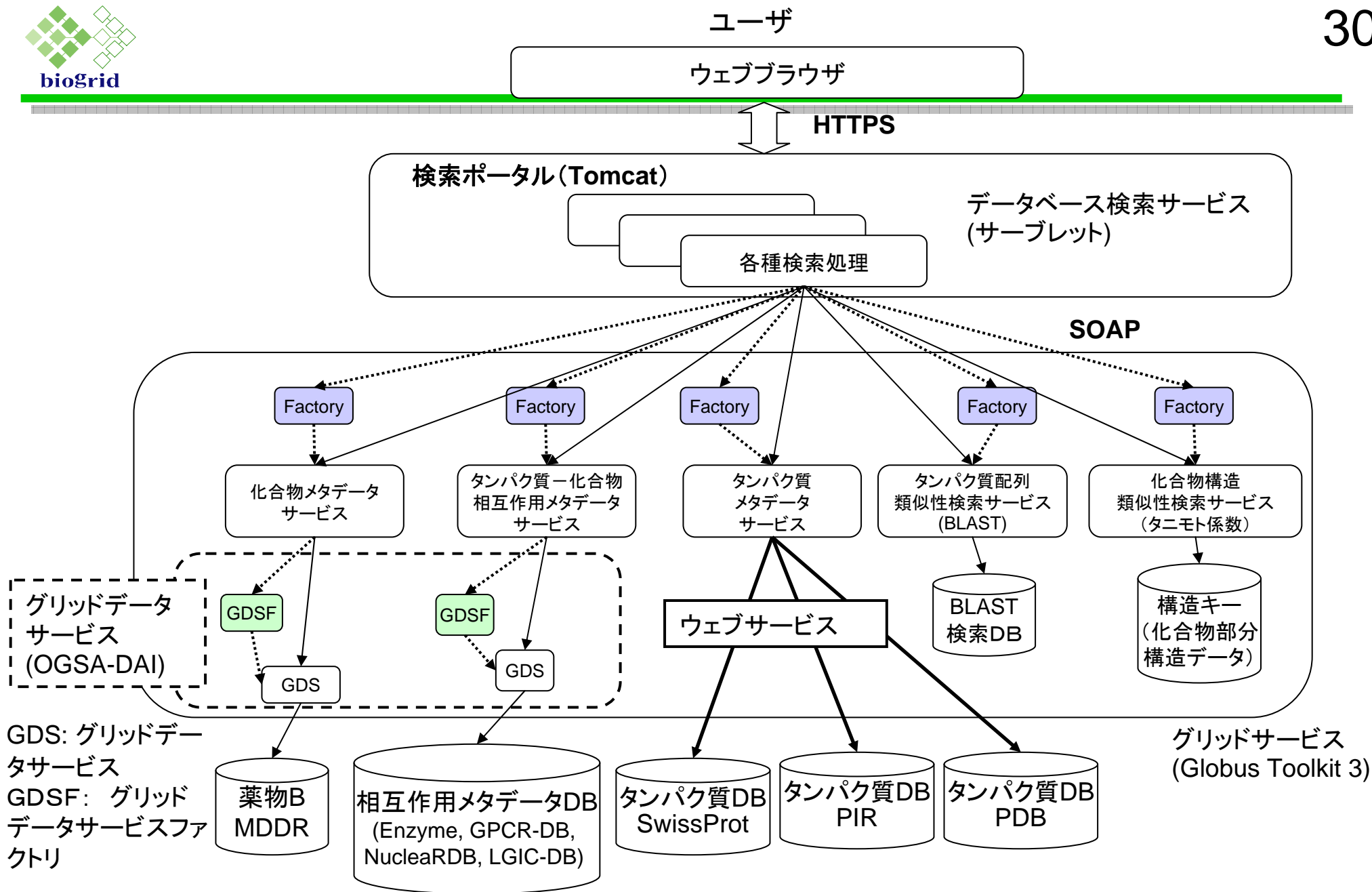
- データベースをデータの種類(分野)ごとにまとめて、統一検索サービスをデータサービスとして実装(OGSA-DAIと類似検索ツールの統合)
- 分野をまたぐデータを利用するアプリケーション(例:ゲノム創薬)のために、データ相互の連携を行うメタデータサービスを実装(独自技術)



OGSA-DAI (Data Access and Integration)

- リレーショナルDBMS(OracleやPostgreSQLなど)やXML-DB(Xindice, eXistなど)をグリッド環境に接続して、GSI認証のもとで利用できる。
- 問合せは、既存のSQLやXPathなどをメッセージに埋め込んで転送する。





現状のデータグリッドシステムのアーキテクチャ

Category	Database	Amount
Disease	Medical Encyclopedia	3079 entries
Genome	DDBJ	Human 7037852 entries, 10176023644 bases Mouse 5063486 entries, 6071844270 bases
Protein	Swiss-Prot	137885 entries, 50735179 amino acids
	PIR	283227 entries, 96134583 amino acids
	PDB	23073 entries
Compound	MDL Drug Data Report (MDDR-3D) Ver. 2003.2	142553 entries
Interaction	Ligand Ontology	ENZYME, GPCR-DB, NucleaRDB, LGIC-DB



DELL PowerEdge 600SC 8台
(Pentium4 2.4GHz,
メモリ 4GB, ディスク400GB)

- 大容量性
 - 1000個以上のDB(毎年約100個のDBが追加)、最大のDB 塩基数200G, エントリ数110M
- 更新頻度
 - 差分データ(毎日)、リリース更新(年4回)、書式更新(年2回)
- メタデータの規格
 - データベースごとにまちまち
- メタデータの共有方法と更新頻度
 - 更新は年数回
 - 乱立するメタデータの相互関連付けが必要
- データポリシ
 - 論文投稿時にデータベース登録が必須
 - 論文採択まで登録のみで、公開を止めておくことが可能
 - 完全に非公開のデータ、有償で利用可能なデータもある

- データの取り扱い
 - 非公開データはネットワークセキュリティで管理(イントラネット、専用線)
 - 有償利用データはSSL (HTTPS)程度のセキュリティ
- データの著作権
 - データそのものには実質的に著作権はない
 - DB登録情報は、DBのIDを明示すれば自由に利用できる
- データにアクセスするコミュニティ
 - 共同研究で複数コミュニティがデータを相互交換してアクセスすることがある
- コミュニティのサイズ
 - グループ単位(数人から数百人までさまざま)でアクセス

- シミュレーションとの関連
 - シミュレーションの初期値や境界条件の設定、結果の検証
 - ドッキングシミュレーションなど、大規模計算機リソースが必要なものがある(共有リソース利用にはセキュリティの問題あり)
 - ワークフローツール(UK e-ScienceのTavernaなど)は潜在的需要は高いが一般的ではない
 - ツールの一般的な共通登録アーカイブはない(コミュニティのレベルではいくつか存在)
- VOの利用
 - グループ単位の利用なのでVOとの親和性が高い
- オープンソース
 - データの更新に伴うプログラムのアップデートや、コミュニティの要求に応じたプログラムのカスタマイズの要求がある
 - ワークフローツールで使うためにも、プログラムのインタフェース公開が必要