DataGrid：グリッド技術によるバイオデータベースの連携

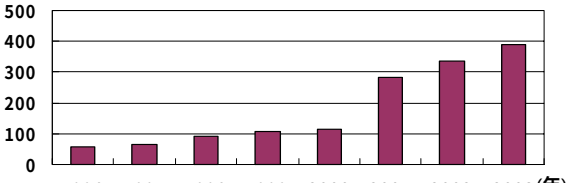
大阪大学　情報科学研究科
松田秀雄

生命科学がグリッドを必要とする理由

- 生命科学はゲノムの登場によってdata driven/intensive scienceに変わりつつある
- 原子、分子、細胞、組織、さまざまなレベルでのシミュレーションが可能であるが、**大規模な計算パワー**を必要とする
- 生命科学は歴史もあり、非常に多岐にわたっているため、そもそも**科学者の協調**が必要である
- 生命科学にまつわる大量の情報が世界中に分散するデータベースで維持されており、それらの**連携**が必要

生命科学におけるデータベース

- ゲノムプロジェクトの成果によるデータベースの数および量の急激な増大



Nucleic Acids ResearchのDB Issueでの論文数

Domain	No. of DBs
DNA	87
RNA	29
Protein	94
Genomic	58
Mapping	29
Protein structure	18
Literature	43
Miscellaneous	153

Current total : 511

GenBankのデータ量の増大
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Bio DB Catalog (DBCAT)への登録数：http://www.infobiogen.fr/services/dbcat/

多数のデータベースの利用法

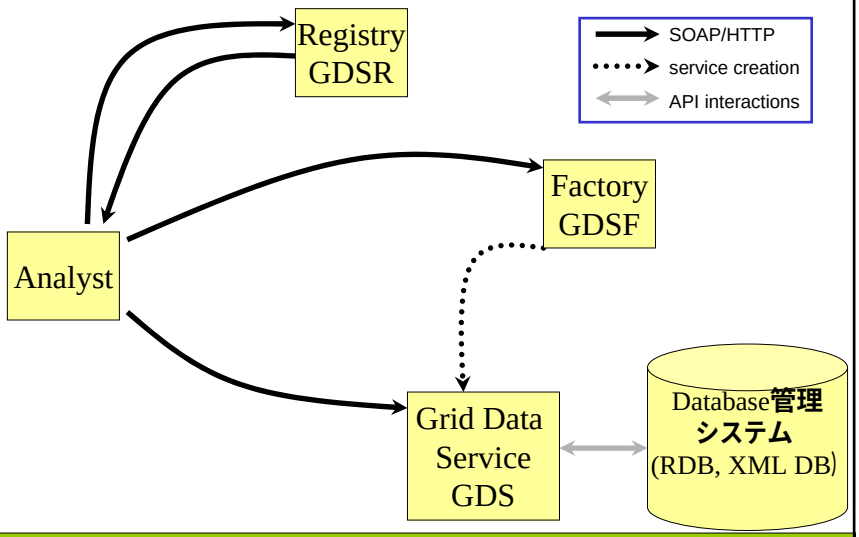
- ハイパーリンク結合（例：京大DBGET／LinkDB）
  - 既に多数のデータベースを連携する方法として実用的に運用されている。
  - ハイパーリンクに「意味」を持たせられない。
- 統合データベース（例：NCBIのEntrez）
  - 個々のデータベースを意識する必要がない。
  - 元のデータベースのスキーマの更新があると、統合のやり直しが必要。
- 異種データベース（例：スタンフォード大学のTSIMMIS）
  - データベースごとにラッパーを作成して共通形式へ動的に変換し、メディエータで相互の検索を仲介（スキーマ更新の影響は1つのラッパーのみ）
  - 生命科学特有のデータベース検索（例えば、配列ホモロジー検索や構造類似検索）と組み合わせることが困難。
  - 商用データベースの利用でのユーザ認証への対応が困難。

→グリッドによるデータベースの連携（データグリッド）

  - データ検索の方法（キーワード検索、全体検索など）ごとにグリッドサービスを提供し、サービス連携によりデータベース間の相互参照を行う。
  - グリッドの認証機構を利用して、各データベースの認証を一元化。

データグリッド基盤システムの開発

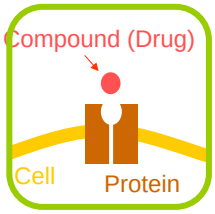
OGSA-DAI（Open Grid Service Architechture Data Access and Integration）の利用

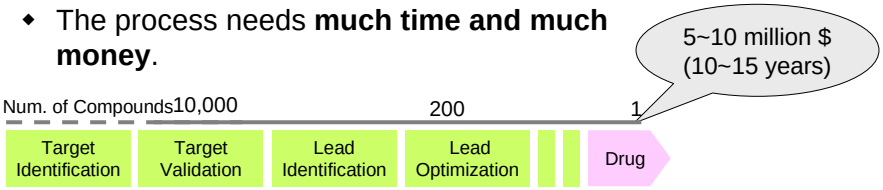


Drug Discovery Process

Application to the **drug discovery process**.

- Compounds (drugs) are activated by binding to proteins in a cell.
- **Drug Discovery Process** is to find chemical compounds that have good effects on their target proteins.
- The process needs **much time and much money**.

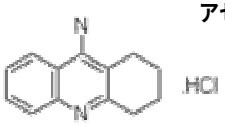




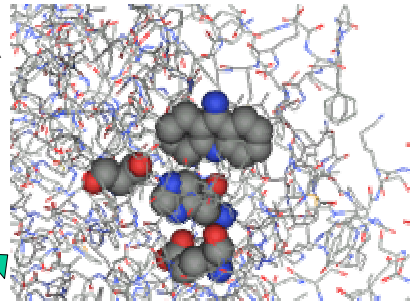
Protein-Compound Interaction Search is one of the most important technologies in drug discovery.
DataGrid can bridge Biology (proteins) and Pharamaceutics (compounds) !

◆データベース連携システムの応用イメージ

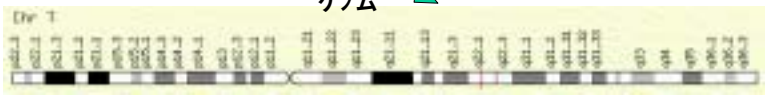
ターゲットタンパク質
アセチルコリンエステラーゼ



Tacrine (Pfizer Cognex)
アルツハイマー病の医薬



ゲノム



SNP (**単一塩基多型**)

Genomic Data								Transcription Data					
SNP ID	Contig Accession	Pos in Contig	Str	5' Flanking Sequence'	3' Flanking Sequence'	Validation	DNA Chg	AA Chg	Type	mRNA Accession	Protein Accession	Cod Pos	Pos in Prot
rs1799806	NT_007933.10	25722157	-	GAGGCTGCTC	GABGCCGGc	by-submitter	G/C	–	retran	NM_000665	NP_000656	–	–
								P/R	coding-nonsynon	NM_015831	NP_056646	2	592
rs7636	NT_007933.10	25723576	-	TGGGGGTGCC	CACGGCTACG	by-frequency	C/T	pgp	coding-synon	NM_000665	NP_000656	3	477

◆ゲノム創薬に必要なデータベース

ゲノム研究	遺伝子探索研究	遺伝子機能解析	ターゲットバリデーション	リード化合物探索	リード化合物最適化	前臨床	臨床	上市

- 疾患モデリング
- ゲノムへのマッピング／遺伝子探索
- 既知タンパク質の検索（配列・構造類似性検索）
- タンパク質・低分子化合物相互作用検索
- 化合物検索（構造類似性など）

◆実施体制

疾患モデリング

ゲノムへのマッピング／遺伝子探索

既知タンパク質の検索（配列・構造類似性検索）

タンパク質・低分子化合物相互作用検索

化合物検索（構造類似性など）

蛋白質研究奨励会

安永（阪大）／HPグループ

藤（京大）／三菱スペースグループ

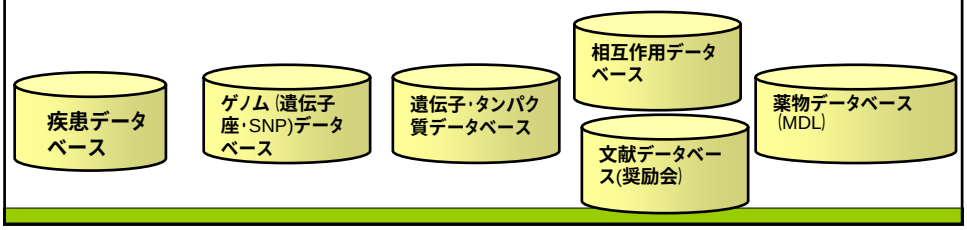
大川／アズテックグループ

松田（阪大）／富士通九州グループ

遠里（阪大）

データグリッド基盤システム

日立ソフト、（基盤グループ　伊達、高坂、上田（阪大））



◆データベース連携システムによる創薬支援の検討

疾患関連情報検索

ゲノムへのマッピング／遺伝子探索

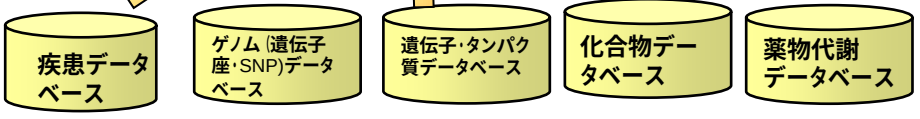
タンパク質検索（配列・構造・機能検索）

タンパク質・低分子化合物相互作用検索

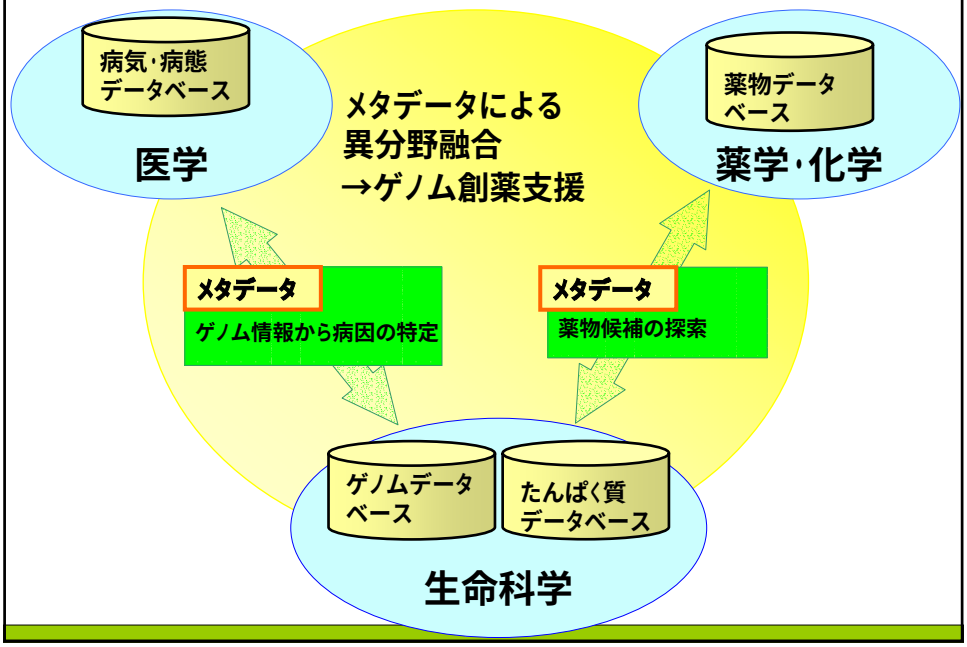
化合物検索（構造類似性など）

アプリケーションとの大きなギャップ

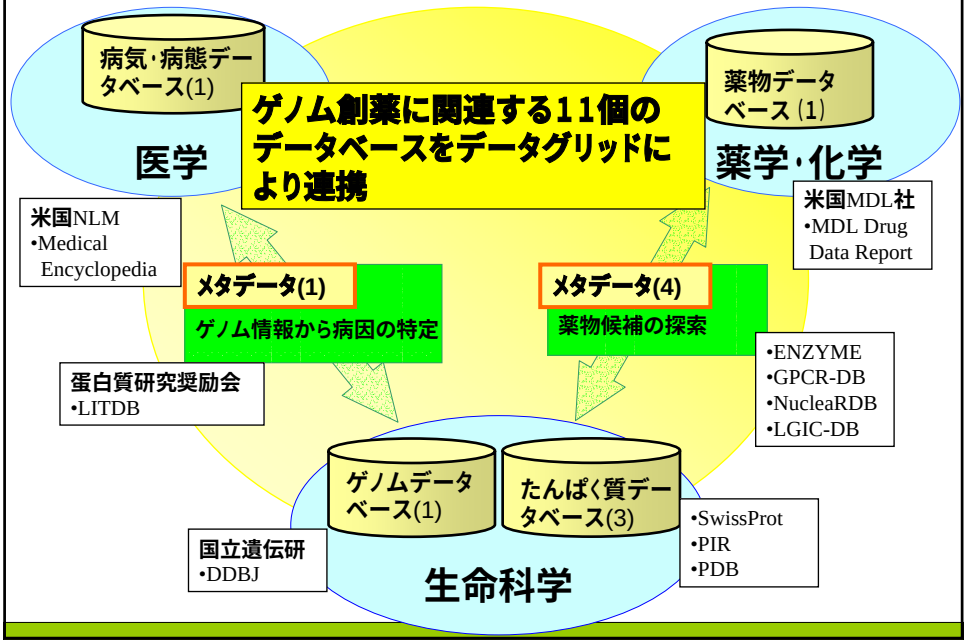
データベース連携システム



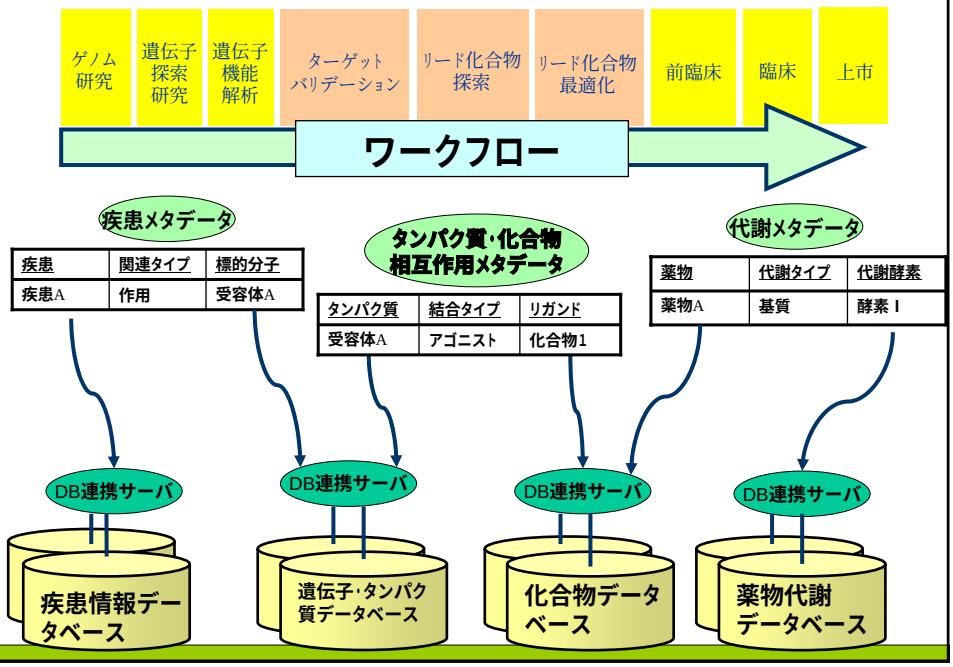
異分野にまたがるデータベース連携



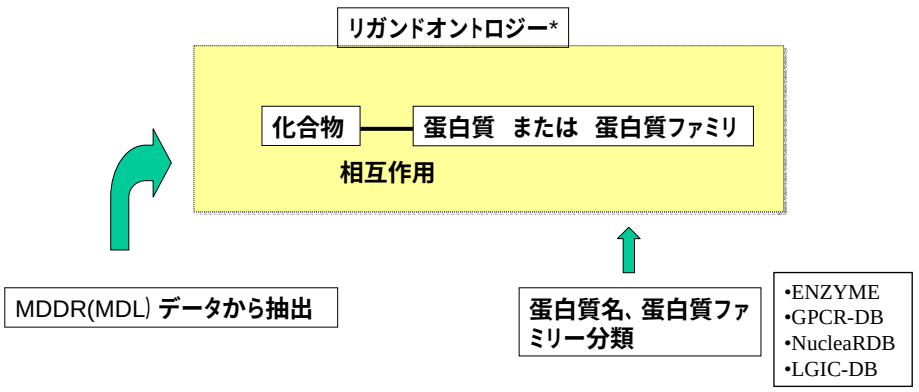
異分野にまたがるデータベース連携



2段階のデータベース連携



タンパク質･化合物相互作用の表現



* Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver JJ, Lecchini S, Jacoby E. "An ontology for pharmaceutical ligands and its application for in silico screening and library design," *J Chem Inf Comput Sci.* 2002 Jul-Aug;42(4):947-55.

類似検索との組合せ：活性相同性検索

活性相同性 (Activity Homology)

新規ターゲットタンパク質

リガンド候補

相同性検索
(BLASTなど)

構造類似性検索
(ISIS SSサーチなど)

化合物デスクリプタ

機能既知のタンパク質 (集合)

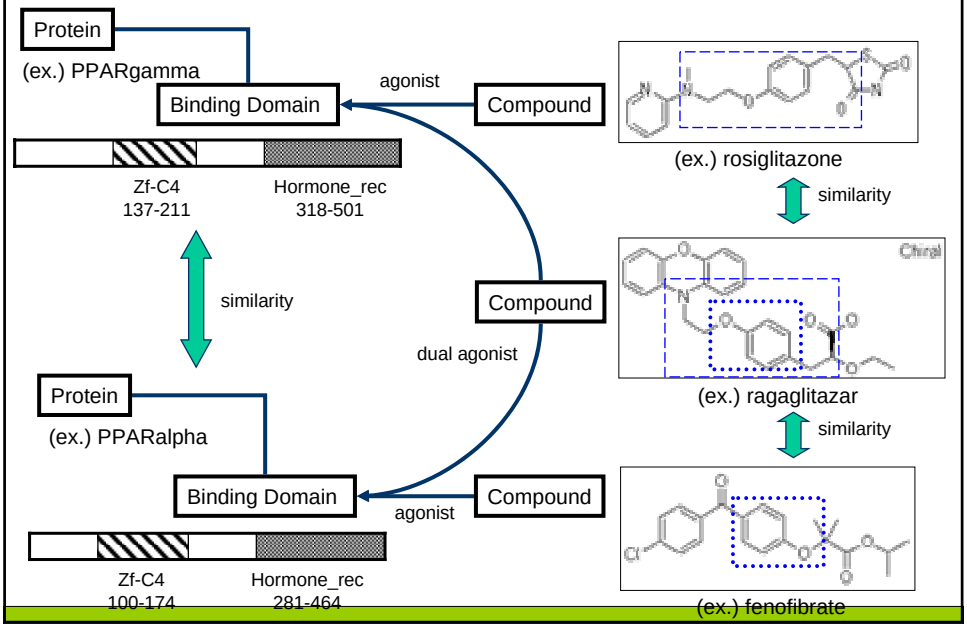
既知のリガンド (集合)

構造解析
(複合体情報、活性部位)

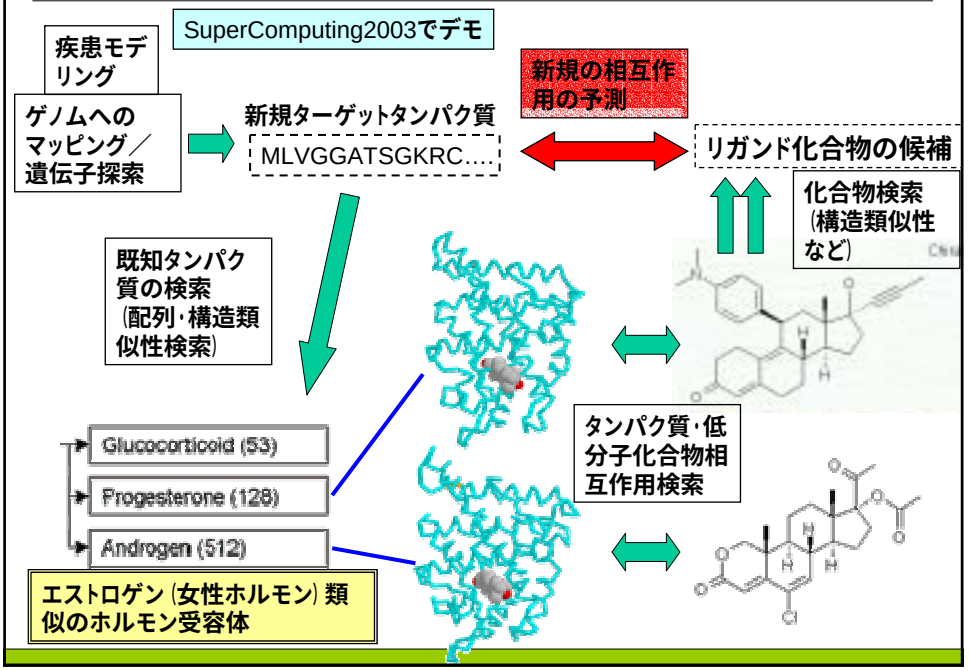
相互作用検索

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.

活性相同性の例

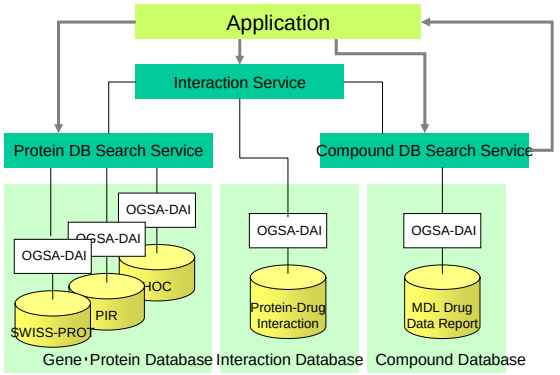


アプリケーションの例



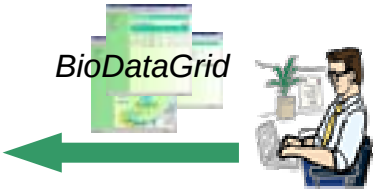
Demonstration System

- Bio-related databases are classified into categories.
- Querying databases for each category are provided as our Grid Service.
- The Grid Services are integrated by using **Globus Toolkit 3** with **OGSA-DAI** (Data Access and Integration)



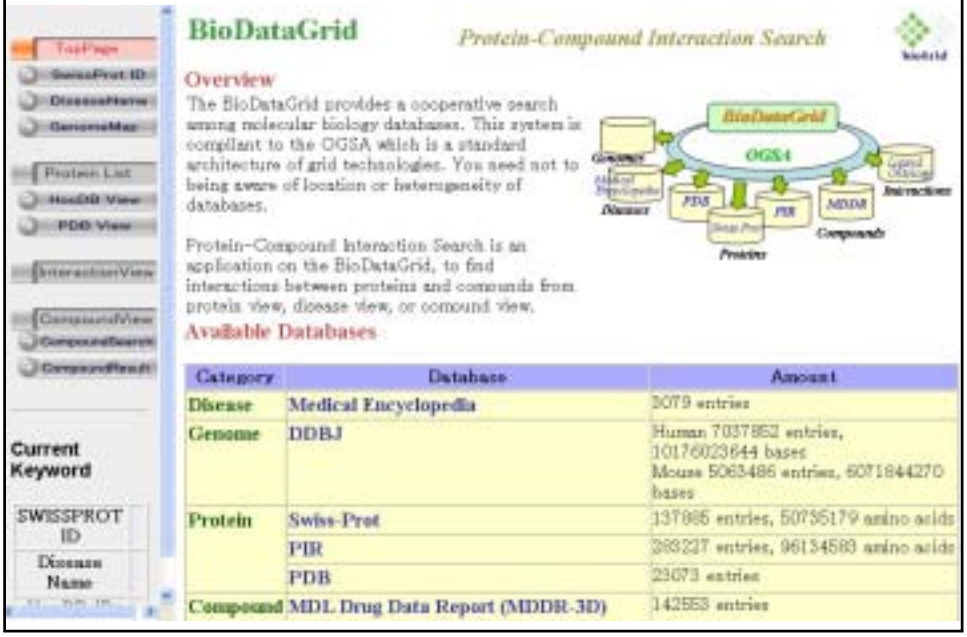
Available databases

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

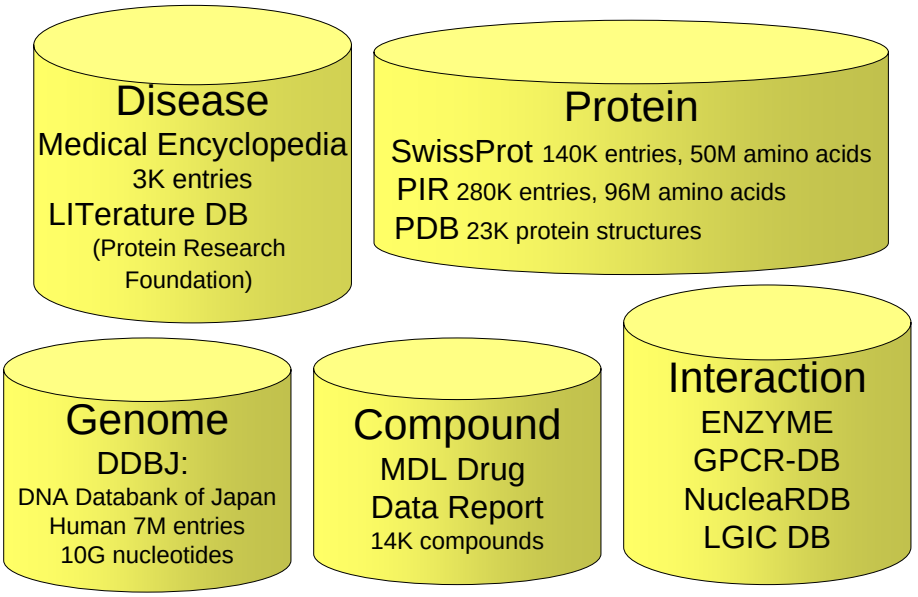


Please come to see our demonstration.

SC2003デモシステム　トップページ



デモシステムで利用したデータベース



デモシステムのソフトウェア構成

1. **Disease View**: Select a disease. Retrieve a protein (target protein) related to the disease.
2. **Genome View**: Check its genome location.

(adjacently-located genes may also be functionally- related)

3. **Homology Search View**: Search for similar proteins against DB.
4. **Protein-Compound Interaction View**: Extract compounds bound to the proteins (using protein-compound interaction metadata).
5. **Compound Search**: Search for new compounds possibly-interacted to the the target protein.

まとめ

- Web上に散在する多数のデータベースを連携するデータグリッド技術を開発。
- ゲノム創薬支援を具体例に、疾患、ゲノム、タンパク質、薬物に関連した11個のデータベースを実際に連携することにより、個別のデータベースを意識させることなく相互のデータを動的に関連付けて検索できるシステムを開発し、SuperComputing 2003でデモを行った。
- 最新のグリッド技術であるGlobus Toolkit 3/OGSA-DAIをいち早く取り入れ、実用的な応用システムを構築することにより、データベース連携のためのデータグリッド技術の有効性を実証。

今後の課題

- **応用面**：
  - メタデータの作成支援（統一的な概念記述、オントロジ）
  - ゲノム創薬の探索過程での統一的なスコアリングによる絞込み（優先度付き探索機構の実現）
- **グリッド基盤技術**：
  - セキュリティ技術の確立
  - 非同期検索(Notification機能）の利用
  - XML DBMS技術の確立