In Silico Discovery

Challenges in Integration and Knowledge Extraction

Su Yun Chung

Research Scientist
San Diego Supercomputer Center
University of California San Diego

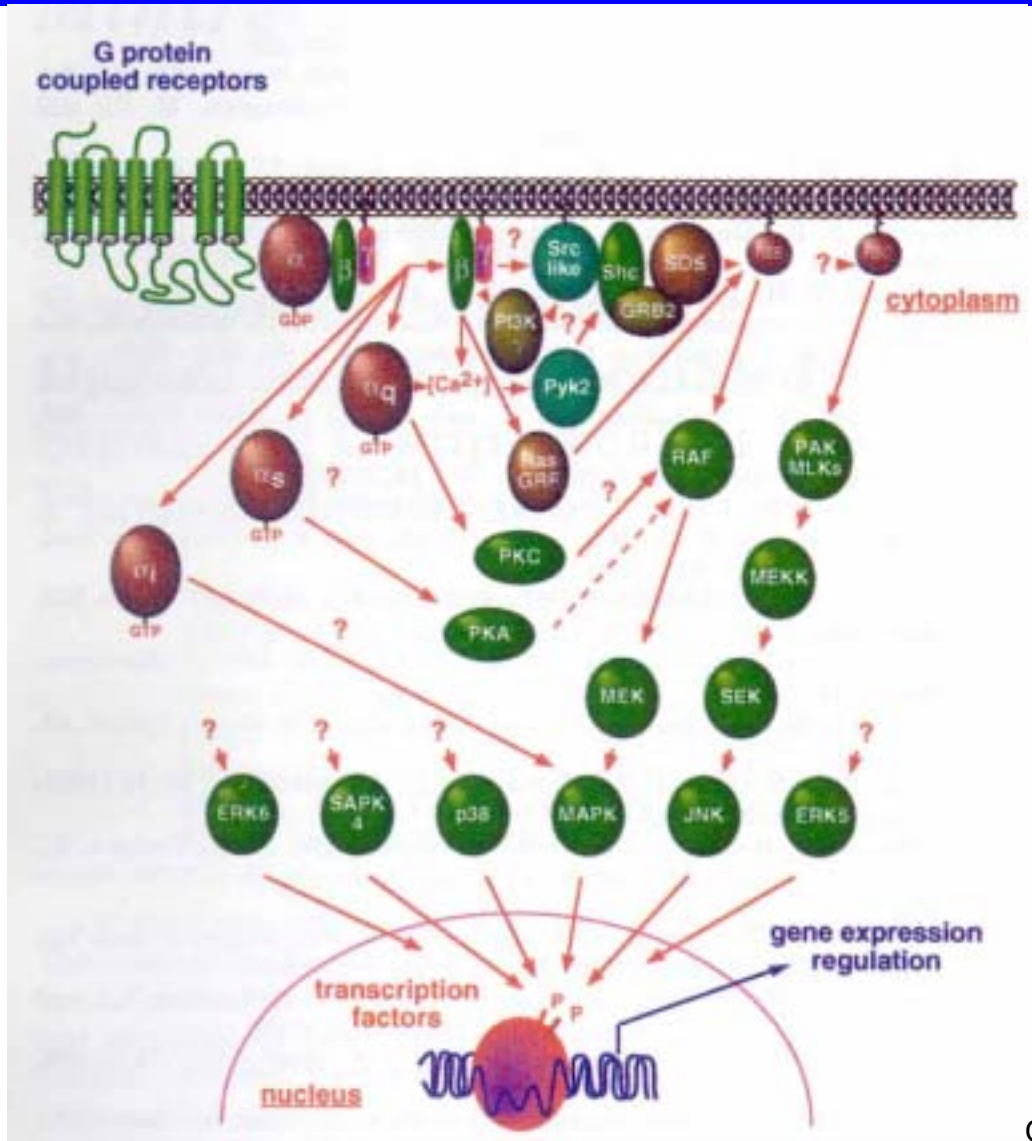
Chief Scientific Officer
geneticXchange, Inc

Taiwan
April, 2002

Outline

- Information Driven Biomedical Research
- The Challenges in Information Integration
- The Solutions
- In Silico Discovery
- Summary

Post-Genomics: World of Proteomics



Gutkind, JBC 273, 1841

Genome-enabled Bioinformatics

- High-throughput technologies generate massive amount of data.

  genome sequencing, microarray gene expression, mass spectroscopy, …

- Growth of data and databases in the public and private domains is ever more rapid.

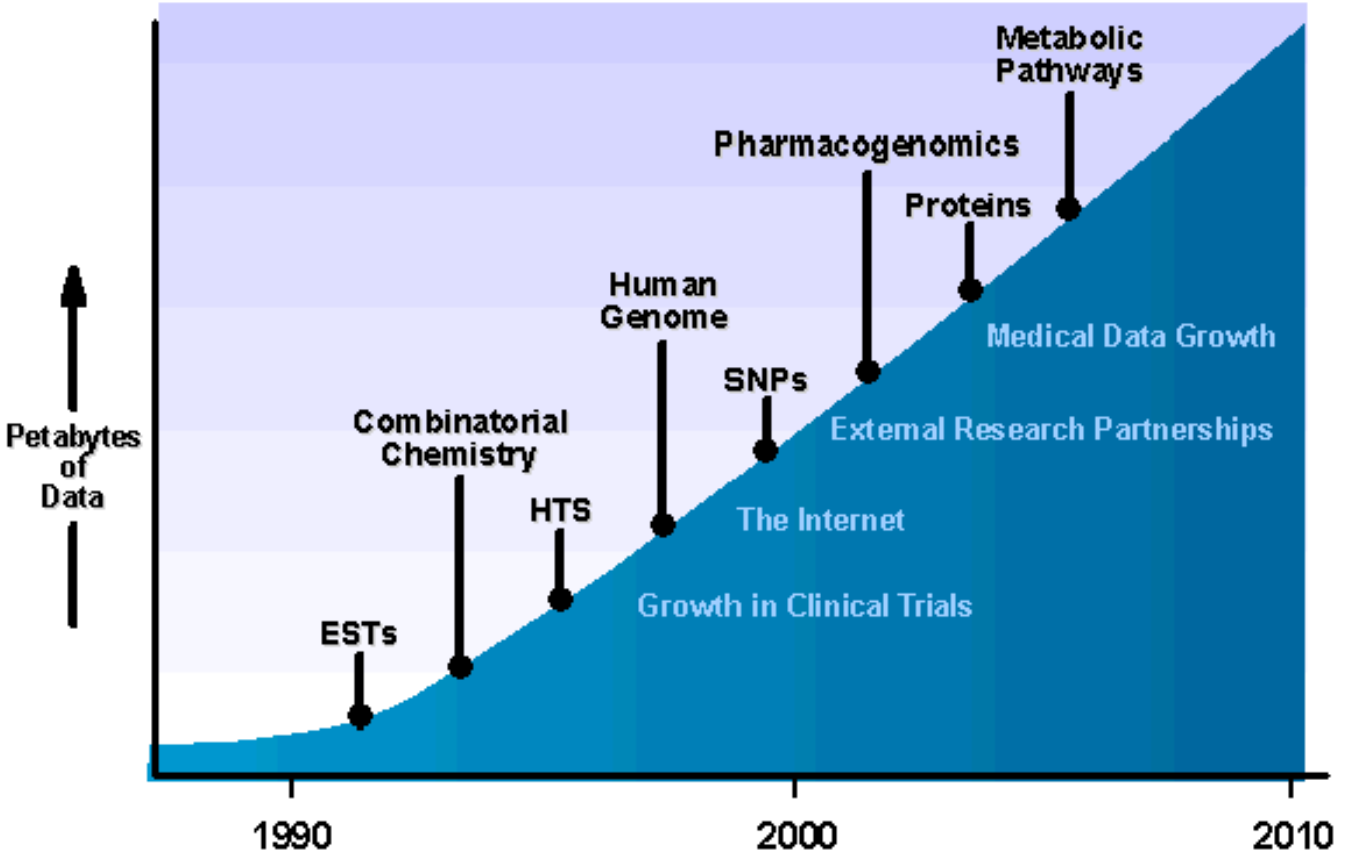
  genomics, gene expression profiles, proteomics, pharmacogenomics, literature, clinical trials…

- Proliferation of computational tools for data analysis and processing continues.

  modeling and simulation, statistical analysis, sequence analysis and gene finding, clustering algorithm, protein folding and structure prediction, data Mining, visualization…

Requirements

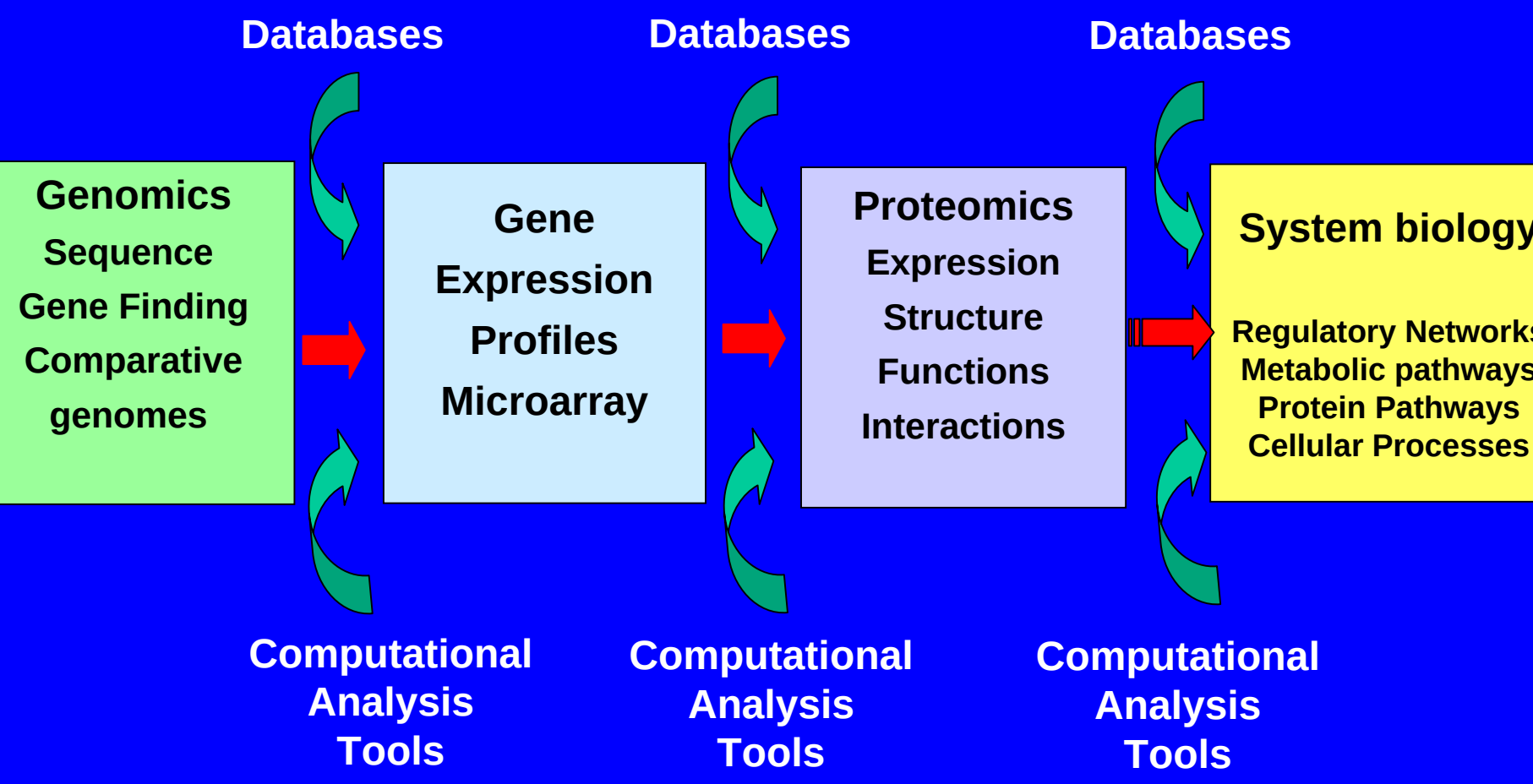
- Life Sciences data is increasing at a tremendous rate
- Petabytes (10^{15}) of data are projected
- Data integration and data management are key to successfully deciphering meaning



The Driving Force

Data ⇨ Information ⇨ Knowledge ⇨ Discovery

Information-Driven Discovery



The Future is Here

- Digitization of biological systems and their processes

  Simulation and modeling of protein-protein interactions, protein pathways, genetic networks, biochemical and cellular processes, normal and disease physiological states,…

- Blurring of the boundary between experimentally generated data and data generated by database searches and computational analyses

- In silico discovery in complement with wet lab experiments

Information Integration

Backbone of Research and Discovery

Understanding brain function requires the integration of information from the level of the gene to the level of behavior. At each of these many and diverse levels there has been an explosion of information, with a concomitant specialization of scientists. The price of this progress and specialization is that it is becoming virtually impossible for any individual researcher to maintain an integrated view of the brain and to relate his or her narrow findings to this whole cloth. Although the amount of information to be integrated far exceeds human limitations, solutions to this problem are available from the advanced technologies of computer and information sciences

NIMH the Human Brain Project Home Page

The biological data and databases

- **Complex and Hierarchical**

  Data types range from sequences, 3-dimensional structures, pathways, images, text, and a wide variety of annotation.

- **Heterogeneous**

  storage format, management, and access vary widely

- **Dynamic**

  contents and schema change routinely and rapidly

- **Inconsistent**

  lack standards at the ontology Level

  - Controlled vocabulary for consistent naming for biomedical terms within and between databases
  - Data models for modeling or abstraction of biological system and processes

What is Ontology?

An ontology is a specification of a conceptualization

Tom Gruber, Computer Science

Ontology provides a vocabulary for representing and communicating knowledge about some topic and a set of relationships that hold among the terms in that vocabulary

Biologists

Life Science Community Efforts

1. Life Science Group/ Object Management Group (OMG)

   http://www.omg.org/lsr/

2. The Human Genome Organization HUGO Gene Nomenclature Committee (HGNC)

   http://www.gene.ucl.ac.uk/nomenclature/

3. Gene Ontology Consortium (GO)

   http://geneontology.org

4. Bio-ontology Consortium : The Molecular Biology Ontology Working Group

   http://smi-web.stanford.edu/projects/bio-ontology

5. The Interoperable Informatics Infrastructure Consortium (I3C)

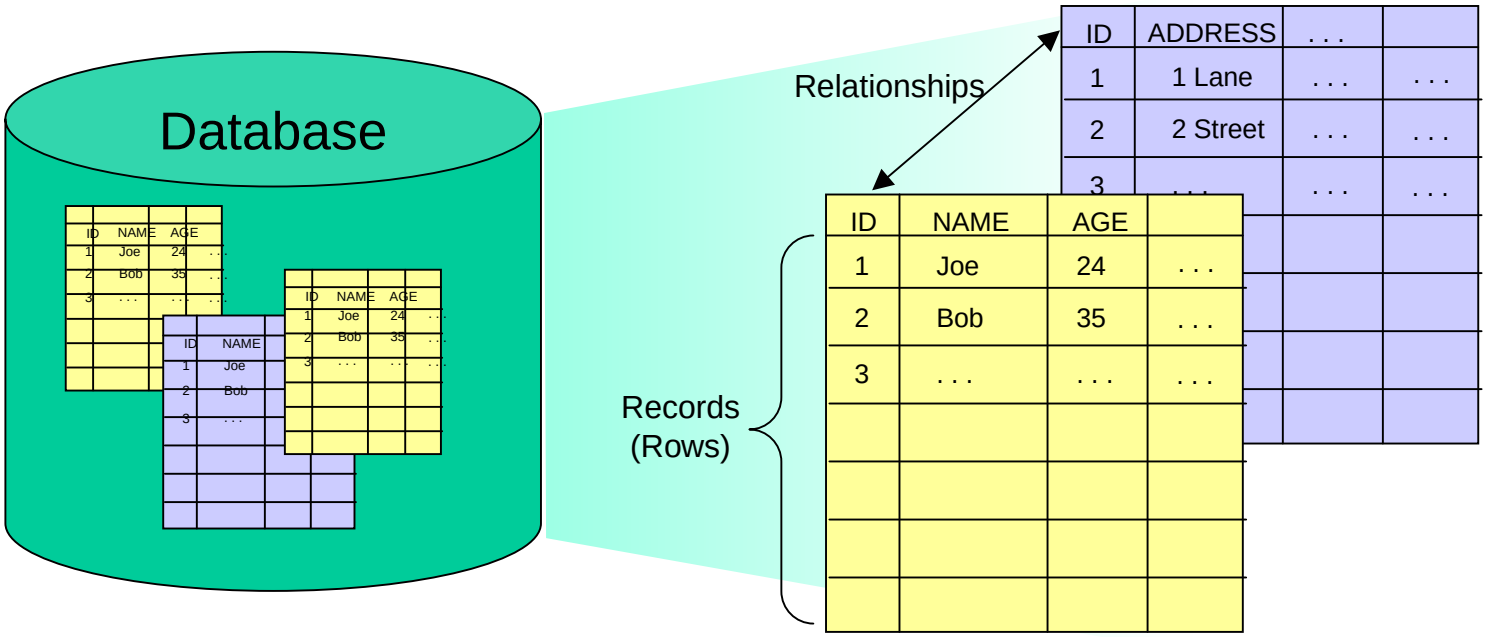
   http://www.i3c.org/

6. SNOMED : Systematized Nomenclature of Medicine

   http://www.snomed.org/

Relational Databases

Tables & Relationships



Semi-Structured

XML

```
<!DOCTYPE memo [
[<!ELEMENT memo   >
<!ELEMENT para    (#PCDATA) >
<!ELEMENT to      (#PCDATA) >
<!ELEMENT from    (#PCDATA) >
<!ELEMENT date    (#PCDATA) >
<!ELEMENT subject (#PCDATA) >]>
 <
<memo>
<to>All staff</to>
<from>Martin Bryan</from>
<date>5th November</date>
<subject>Cats and Dogs</subject>
<text>Please remember to keep all
cats and dogs indoors tonight.</text>
</memo>
```

- Exchange format
- Nested data structures

Flat File

ASN.1

```
LOCUS       AF169225_1    413 aa                PRI       14-MAY-2001
DEFINITION  beta-2-adrenergic receptor [Homo sapiens].
ACCESSION   AAD48036
PID         g5714688
VERSION     AAD48036.1  GI:5714688
KEYWORDS
SOURCE      human.
      ORGANISM  Homo sapiens
                Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
                Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 413)
AUTHORS   Rupert,J.L., Monsalve,M.V., Devine,D.V. and Hochachka,P.W.
TITLE     Beta2-adrenergic receptor allele frequencies in the Quechua, a high altitude native population
JOURNAL   Ann. Hum. Genet. 64 (2), 135-143 (2000)
MEDLINE   21141798
PUBMED   11246467
FEATURES             Location/Qualifiers
     source          1..413
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="5"
                     /map="5q31-q33"
                     /cell_type="lymphocyte"
                     /tissue_type="blood"
                     /note="isolated from a Quechua speaking native American  heterozygous for a known
     Protein         1..413
                     /product="beta-2-adrenergic receptor"
     CDS             1..413
                     /coded_by="AF169225.1:17..1258"
ORIGIN               1 mgqpgngsaf llapngshap dhdvtqqrde vwvvgmgivm slivlaivfg nvlvitaiak
                    61 ferlqtvtny fitslacadl vmglavvpfg aahilmkmwt fgnfwcefwt sidvlcvtas
                   121 ietlcviavd ryfaitspfk yqslltknka rviilmvwiv sglxsflpiq mhwyrathqe
                   181 aincyanetc cdfftnqaya iassivsfyv plvimvfvys rvfqeakrql qkidksegrf
                   241 hvqnlsqveq dgrtghglrr sskfclkehk alktlgiimg tftlcwlpff ivnivhviqd
                   301 nlirkevyil lnwigyvnsg fnpliycrsp dfriafqell clrrsslkay gngyssngnt
                    361 geqsgyhveq ekenkllced lpgtedfvgh qgtvpsdnid sqgrncstnd sll
```

Structured

Nested data

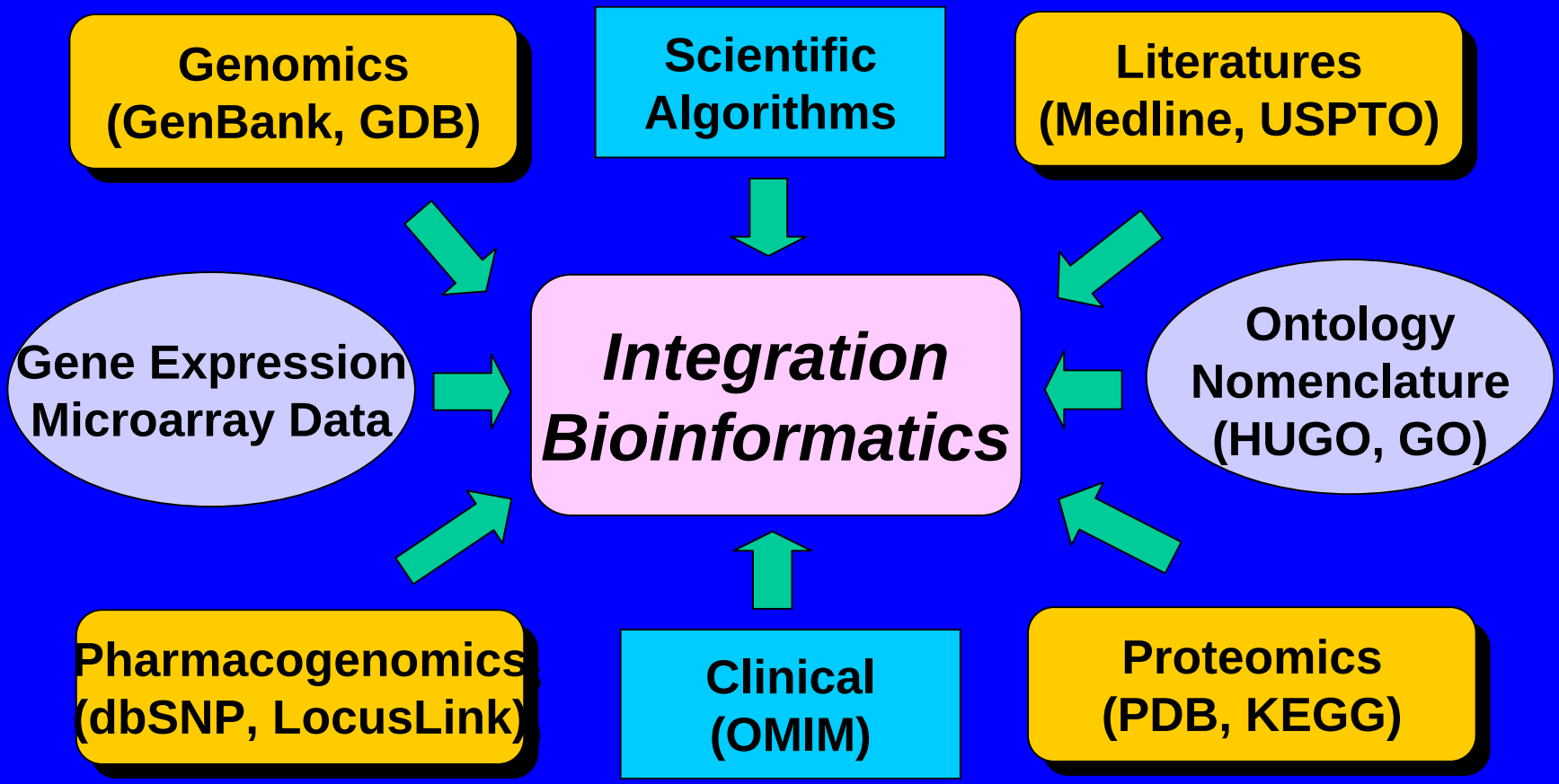
Multiple data types

Alias for a Transcription Factor

CEBPB (HUGO Gene Symbol)

- CCAAT/ENHANCER-BINDING PROTEIN, BETA
- C/EBP-BETA
- CRP2
- INTERLEUKIN 6-DEPENDENT DNA-BINDING PROTEIN
- IL6DBP
- NFIL6
- LIVER ACTIVATOR PROTEIN
- LAP
- LIVER-ENRICHED TRANSCRIPTIONAL ACTIVATOR PROTEIN
- TRANSCRIPTION FACTOR 5
- TCF5

Challenges: Integrate databases & algorithms

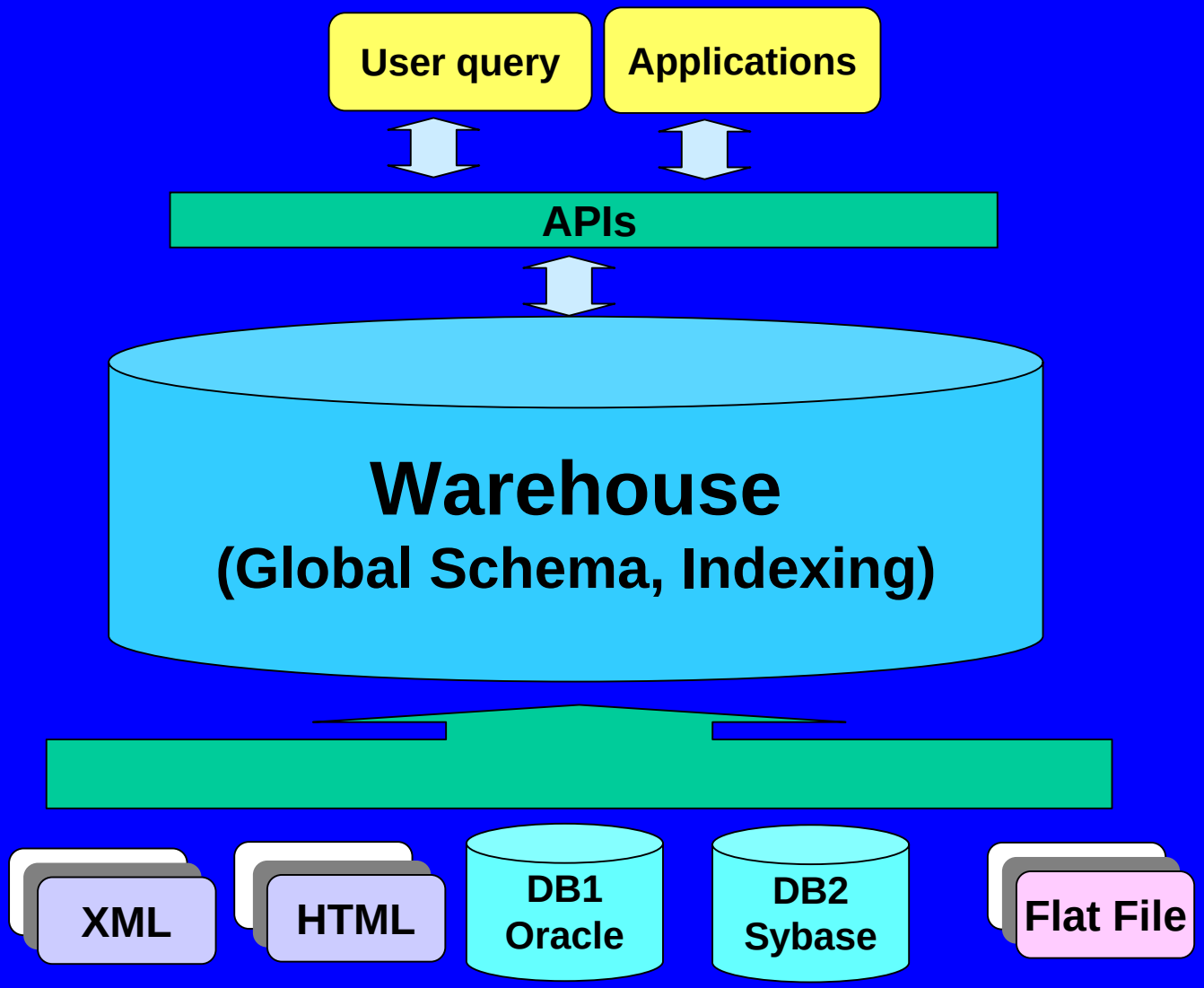


What are the solutions?

- Hypertext navigation by point and click
- The IT (Information Technology) approach by hard coding

  (Perl scripts, C++, JAVA,…)
- The data warehousing approach
- The mediation approach

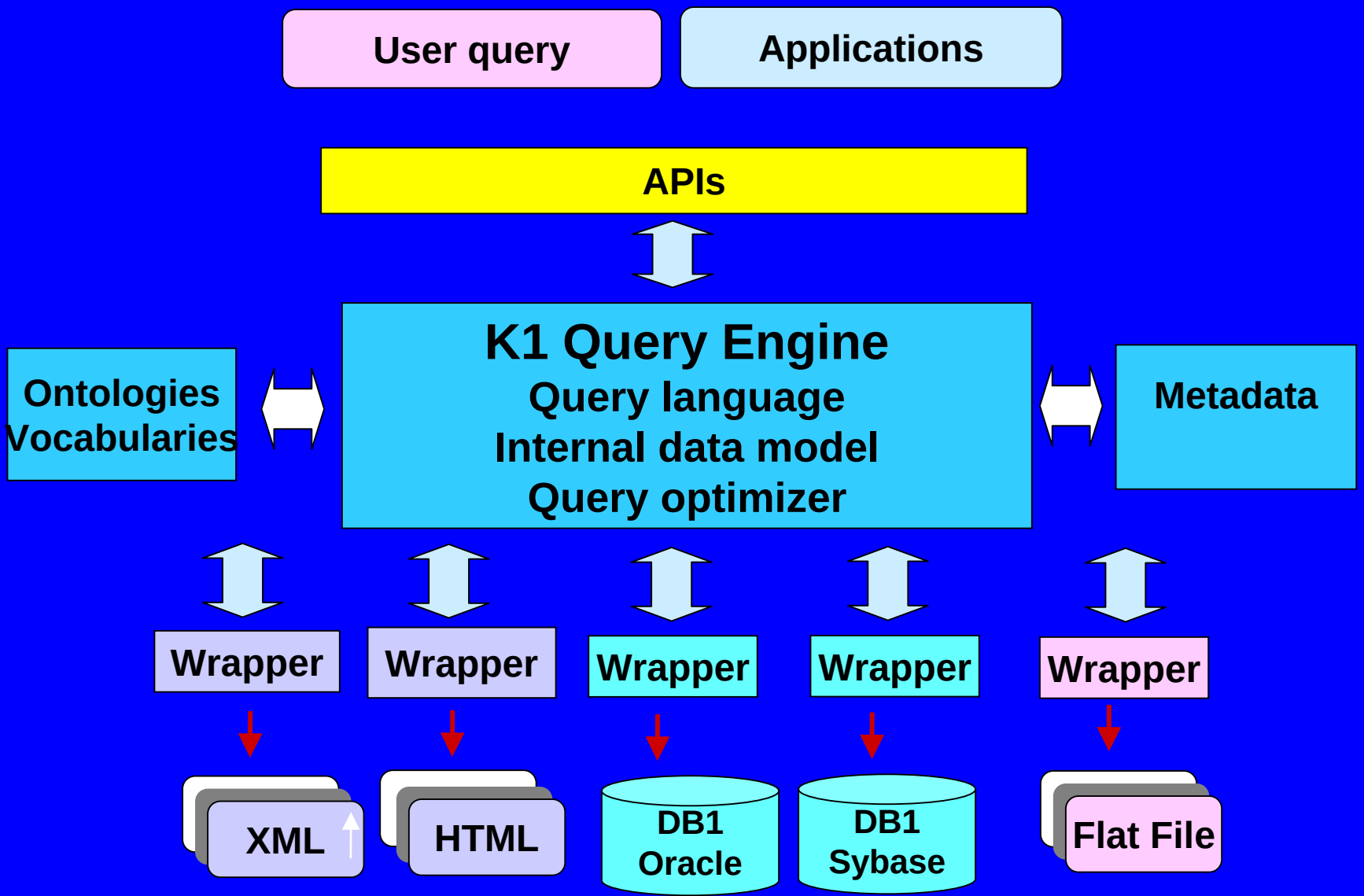
The Data Warehousing Approach



Data Warehousing

- High demand on storage capacity
- Unable to keep up with rapid changes in data content and data schema of sources
- High cost of maintenance

The Solutions: Mediation Approach



The K1 query engine

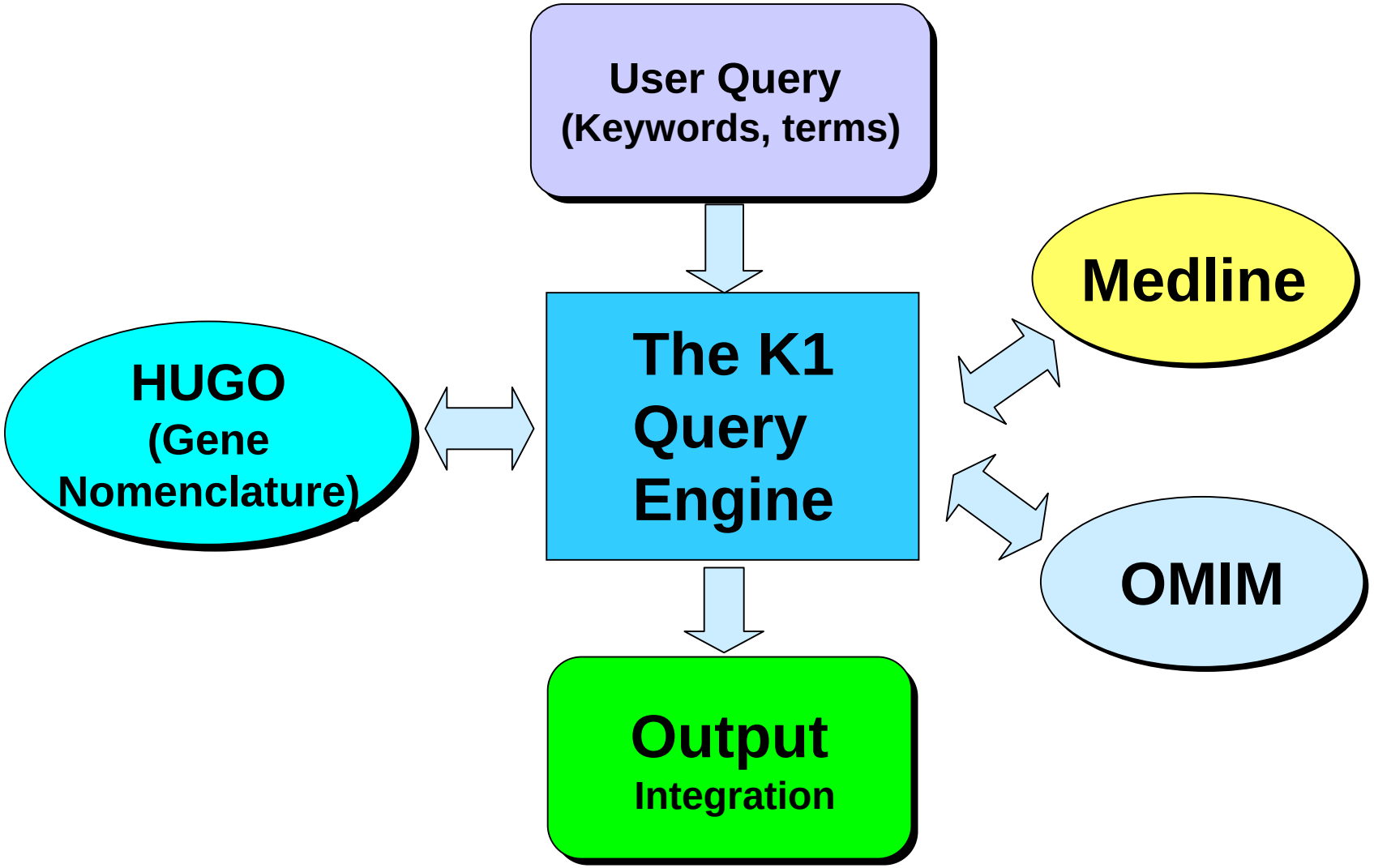
- A federated approach in data integration
- A powerful high-level query language to transform, manipulate, and integrate data
  - SQL-like syntax
- An internal, nested complex data model that facilitates data exchange and data integration
  - The data model encompasses existing popular data formats: XML, HTML, commercial relational data models (RBDMS), flat files, etc.
- A robust query optimizer
  - Parallel processing and multi-threading

Samples of geneticXchange wrappers

- DBMS
  - Oracle
  - Sybase
  - DB/2
  - Informix
  - MySQL
- Algorithms
  - BLAST
  - FASTA
  - ClustalW
  - HMMER
  - BLOCKS
  - Pfscan
  - nnPredict
  - PSORT

- Bio Data sources
  - GenBank
  - Entrez/NCBI
  - SwissProt
  - Locus Link
  - UniGene
  - dbSNP
  - OMIN
  - Taxonomy
  - PDB
  - SCOP
  - TIGR
  - Medline
  - KEGG
  - USPTO

How does it work? An example



Query Multiple Databases with Ontology

- HUGO (Gene Nomenclature Genew DB)
- OMIM (Online Mendel Ian Inheritance of Man)
- MEDLINE

```
Select
(Hugo: x,
  OMIM:  omim-get-detail (x.MIM),
  PMID1_ABS: ml-get-abstract-by-uid (x.PMID1),
  NUM_Aliases: ml-get-count-general (x.Aliases)))
  from hugo-get-ids() x
  where x.Symbol = "CEBPB";
```

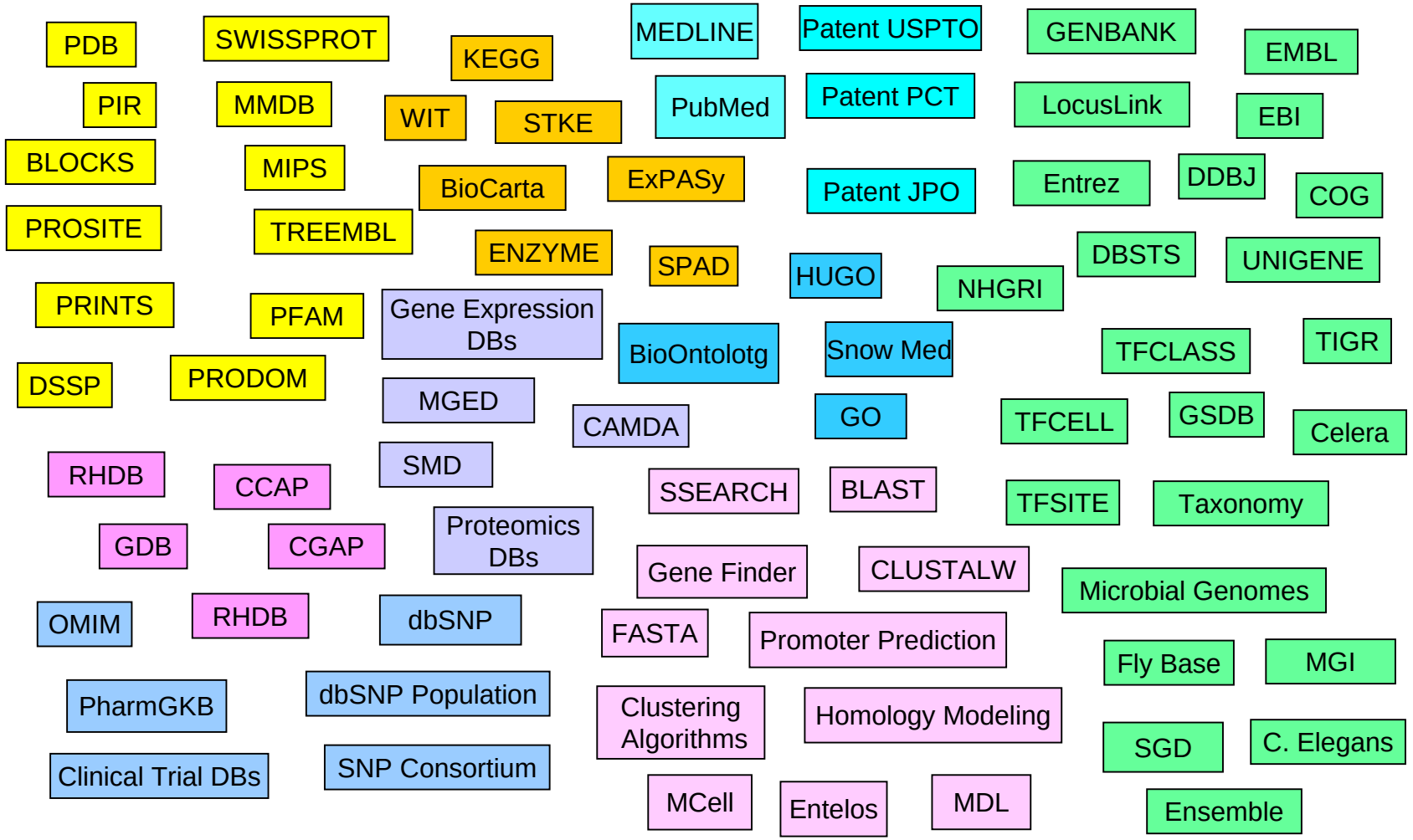
Query Results

```
{(#HGNC: "1834",  #Symbol: "CEBPB",
  #Name: "CCAAT/enhancer binding protein (C/EBP), beta",
  #MIM: "189965",  #PMID1: "1535333",
  #Aliases: "LAP, CRP2, NFIL6, IL6DBP“)
 #OMIM: {(#uid: 189965,
              #gene_map_locus: "20q13.1",….
              #allelic_variants: {})},
  #PMID1_ABS: {(#muid: 1535333,
            #authors: "Szpirer C,…",
            #address: "Departement de Biologie…",
            #title: "genes encoding the liver-enriched
                        transcription factors C/EBP,…",
            #abstract: "By means of somatic cell hybrids
              segregating either human…..”
                  #journal: "Genes Dev  1991 Sep;5(9):1538-
                                                    52")},
  #NUM_entries: 1936)}
```

Advantages

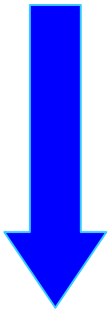
- On Demand Access
  - **The most up-to-date, relevant data sources**
  - **The best-of-breed computational tools**
- Real-time information integration for rapid prototyping and decision-making support
- Flexible transformation and manipulation of data
- The right **information** in the right **context**
- In Silico discovery

Swimming in Data Sources



What's in it for the Biologists?

Information Integration



In Silico Discovery Kits (ISDKs)

What is an In Silico Discovery Kit (ISDK)?

An in silico discovery kit is a script written in the query language that

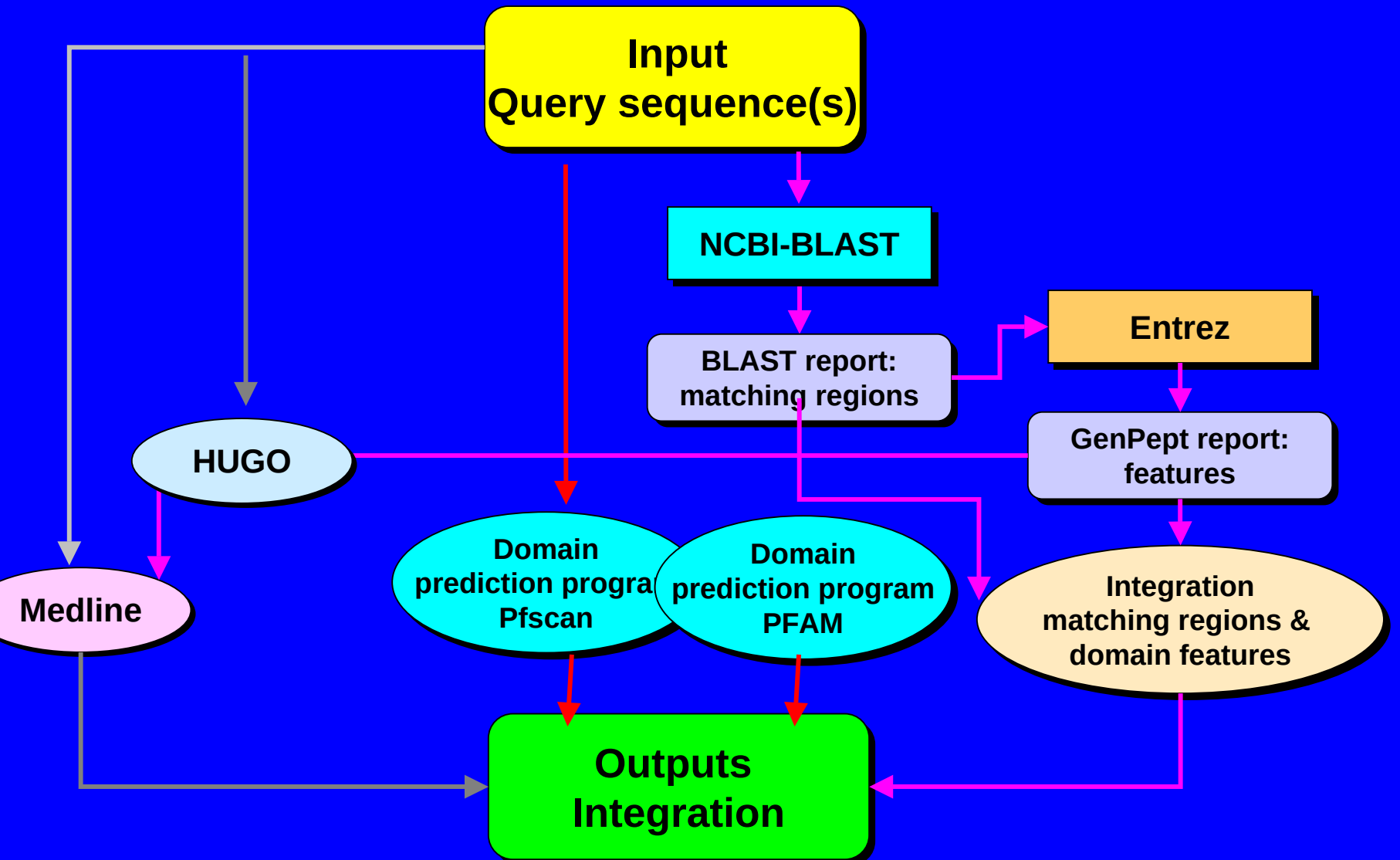
1. **inputs user data and parameters**
2. **performs a defined information integration task**
3. **output the results**

The In Silico Discovery Kit (ISDK)

- Connects multiple computational steps of database queries and applications of analytical tools or algorithms very much like an integrated circuitry
- Pipelines data flow from one step to the next seamlessly
- Execute automatically through the K1 engine
- Runs in batch mood for high-throughput data processing

A Sample of In Silico Discovery Kit (ISDK)

Protein functional domain annotation



ISDKs: the building blocks of discovery

Like Lego-blocks, simple ISDKs can be used to build more sophisticated discovery processes. For example, ISDKs for gene expression analysis, protein functional domain prediction, SNP analysis, and clinical trials can be chained together to form a target identification kit for drug discovery.

ISDKs

- Flexible

  The modular approach of ISDK gives scientists the flexibility to select and combine specific ISDKs for specific research project.

- Scalable: high throughput bioinformatics processing

  The ISDKs are executed automatically by the powerful gX system in batch mode and can handle high throughput data processing

- Re-usable codes

  The ISDK scripts are reusable to perform repetitive tasks and can be shared among scientific collaborators

- Customizable

  ISDKs provide a base set of templates for bioinformatics integration. Thes templates can be readily modified to meet user needs

- Updateable

  New databases and new algorithms or computational tools can be readily

Summary

- A dynamic federated database approach in data integration
- A powerful information integration approach
- A plug and play technology that provides non-intrusive enhancement of existing bioinformatics infrastructure
- Innovative in silico discovery kits (ISDKs) that improve the efficiency and productivity of research in the life sciences

contact info:

suchung@sdsc.edu