



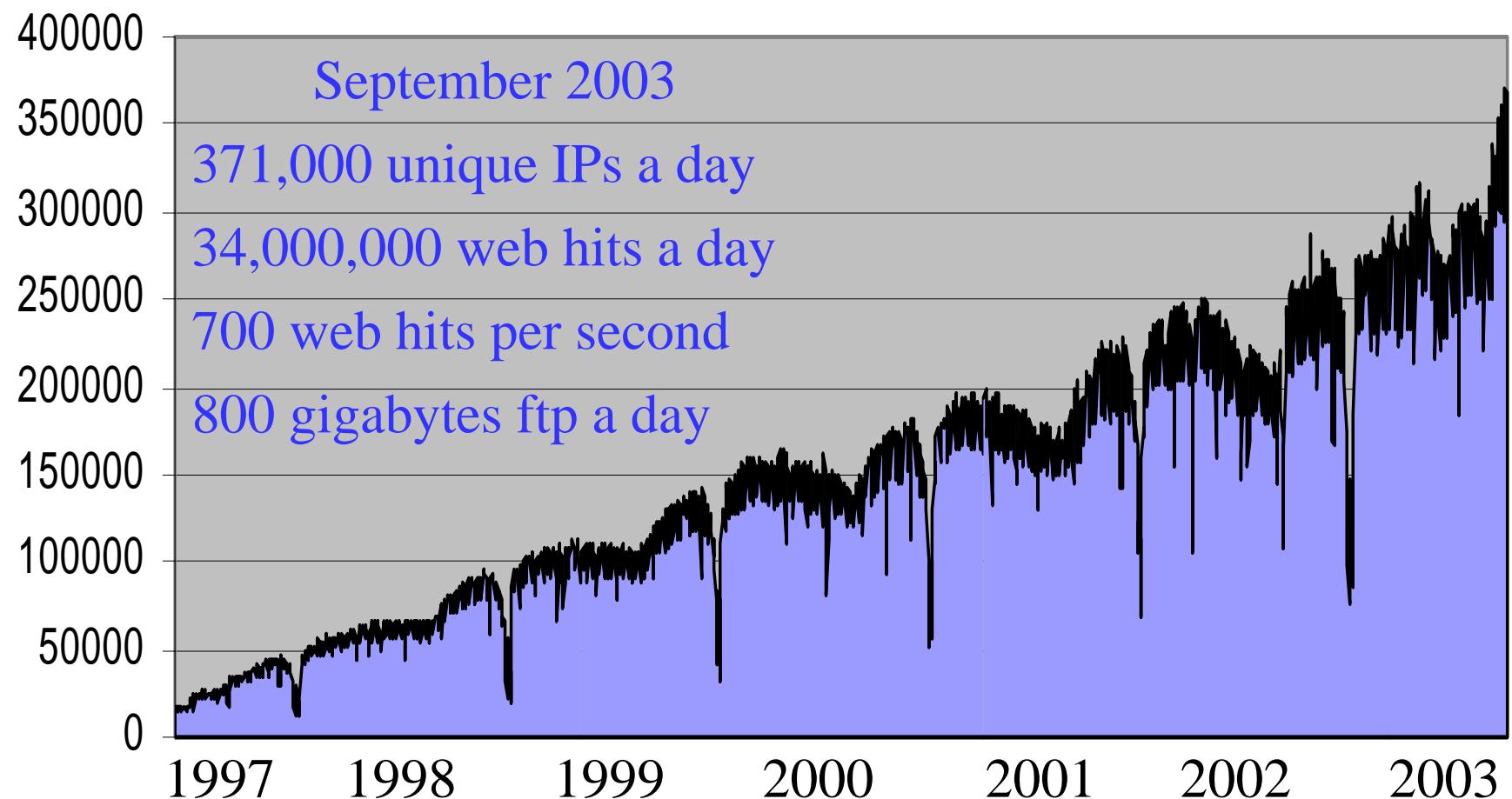
# National Center for Biotechnology Information

- Created by Public Law 100-607 in 1988 as part of National Library of Medicine at NIH to:
  - Create automated systems for knowledge about molecular biology, biochemistry, and genetics.
  - Perform research into advanced methods of analyzing and interpreting molecular biology data.
  - Enable biotechnology researchers and medical care personnel to use the systems and methods developed.
- Builders and providers of GenBank, Entrez, Blast, PubMed, UniGene, OMIM, LocusLink, RefSeq. Online systems host more than 2 million users per month.
- Center for basic research and training in computational biology.

NCBI

WWW  
110101

## Unique Users Per Day

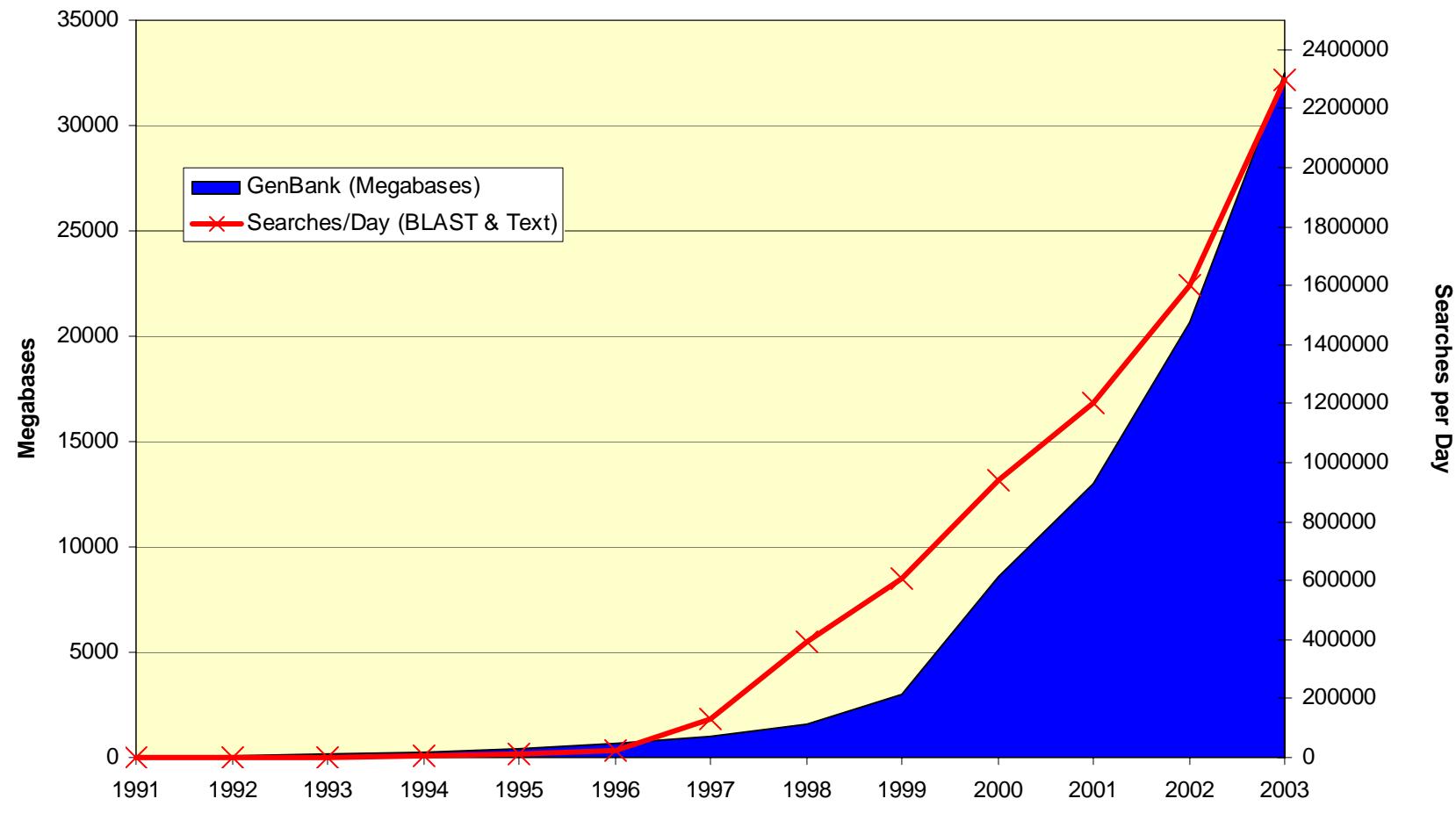


NCBI

WWW 110101

# More Data

Growth of Searches and GenBank



NCBI

WWW  
110101

# Comparative Analysis in Biology



Human

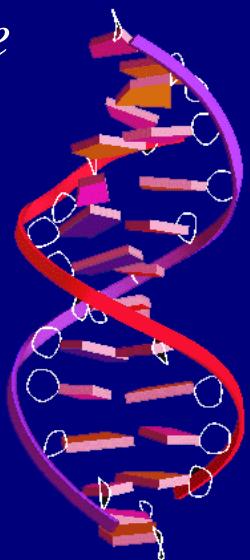


Dog

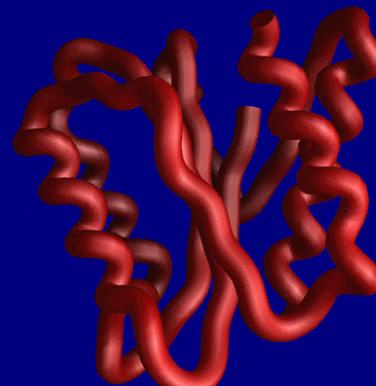


# Genotype to Phenotype

*Gene*



*Function*



> DNA sequence

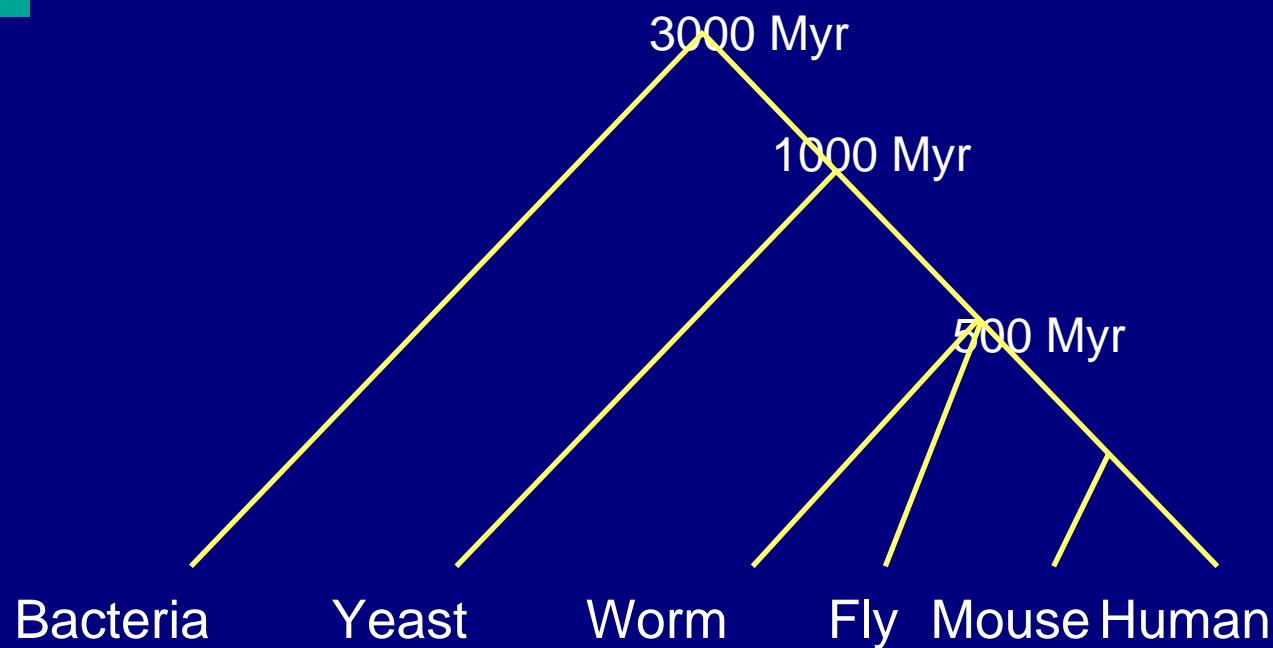
```
AATT CATGAAAATCGTATACTGGTCTGGTACCGCAACAC  
TGAGAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA  
TCTGGTAAAGACGTCAACACCATCAACGTGTGACGTTA  
ACATCGATGAACTGCTGAACGAAGATATCCTGATCCTGGG  
TTGCTCTGCCATGGCGATGAAGTTCTCGAGGAAAGCGAA  
TTTGAACCGTTCATCGAACGAGATCTCTACCAAAATCTCTG  
GTAAGAAGGTTGCGCTGTTGGTTCTTACGGTTGGGGCGA  
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC  
TACGGTTGC GTTGGT GAGACCCCGCTGATCGTT CAGA  
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAATTGG  
TAAGAAGATCGCGAACATCTAGTAGA
```

> Protein sequence

```
MKIVYWSGTGNTEKMAELIAKGIIIESGKDVTNTINVSDVNI  
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK  
KVALFGSYGWGDGKWMRDFFERMNGYGCVVVETPLIVQNE  
PDEAEQDCIEFGKKIANI
```



# Comparative Analysis of Genes



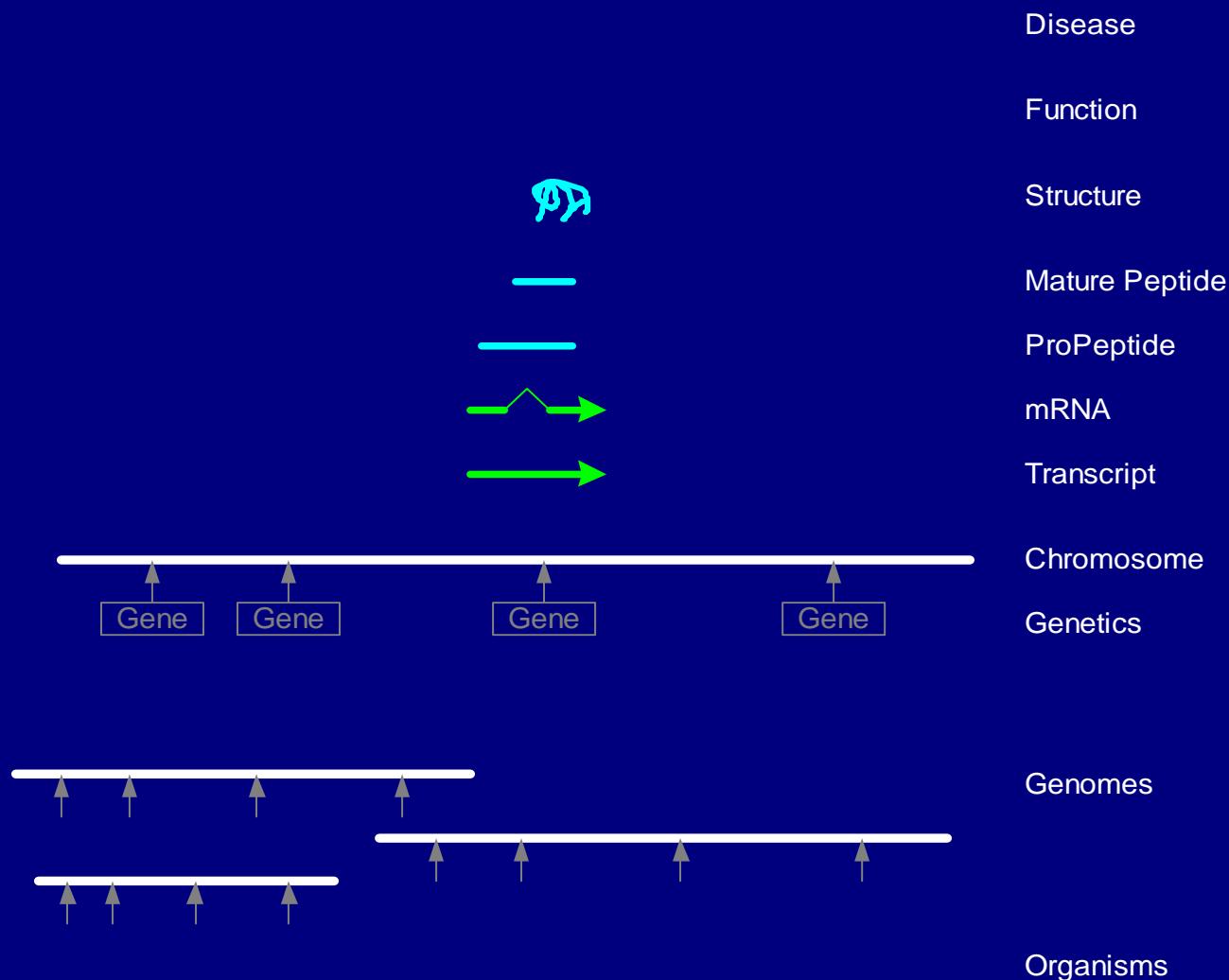
Human 638 RHACVEVQDEIAF**I**PNDVYFEKDKQMFH**I****I**TGPNMG**K**STY**I**R**Q**TGVIVL**M**A**Q****I**GCF**V**PC 697  
Yeast 657 RHPVLEM**Q**DDIS**F**ISNDVTLES**G**K**D**FL**I****I**TGPNMG**K**STY**I**R**Q**VG**V****I**S**L****M****A****Q****I**GCF**V**PC 716  
*E. coli* 584 RHPVVEQVLNEP**F**IANPLNLSPQR**R**-ML**I****I**TGPNMG**K**STY**M**R**Q**TAL**I****A****L****M****A****Y****I****G****S****Y****V**PA 642

Colon cancer gene sequence

NCBI

110101

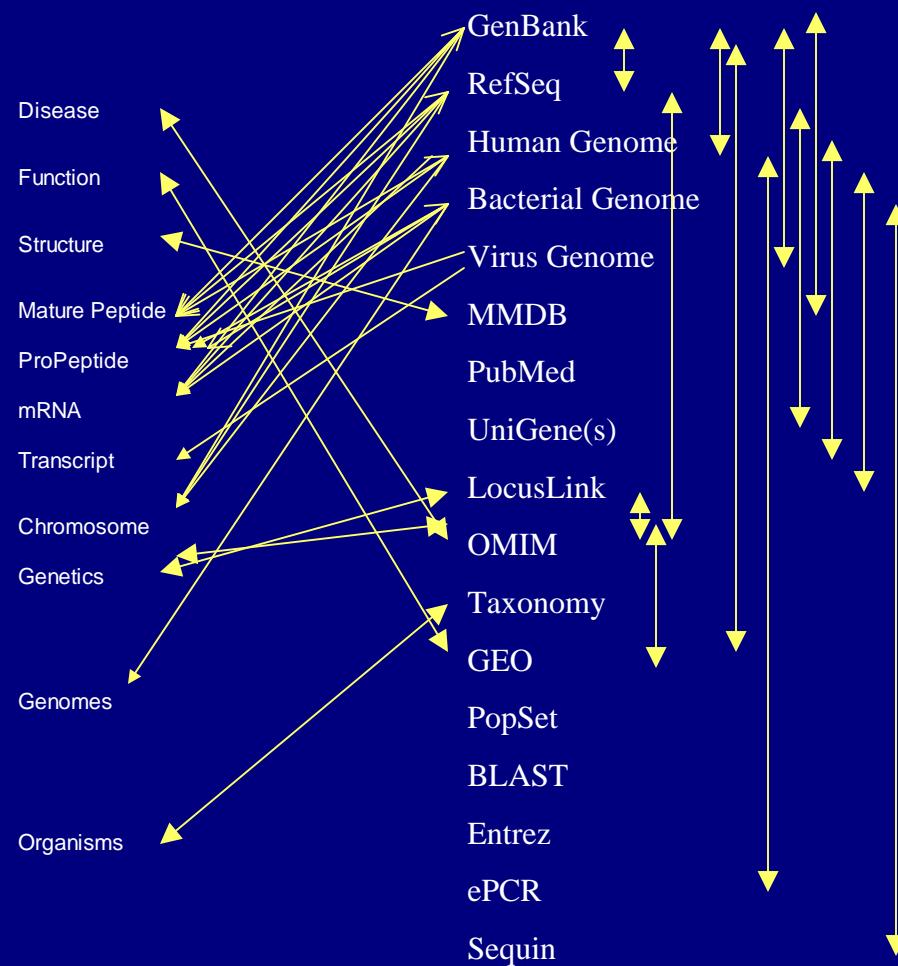
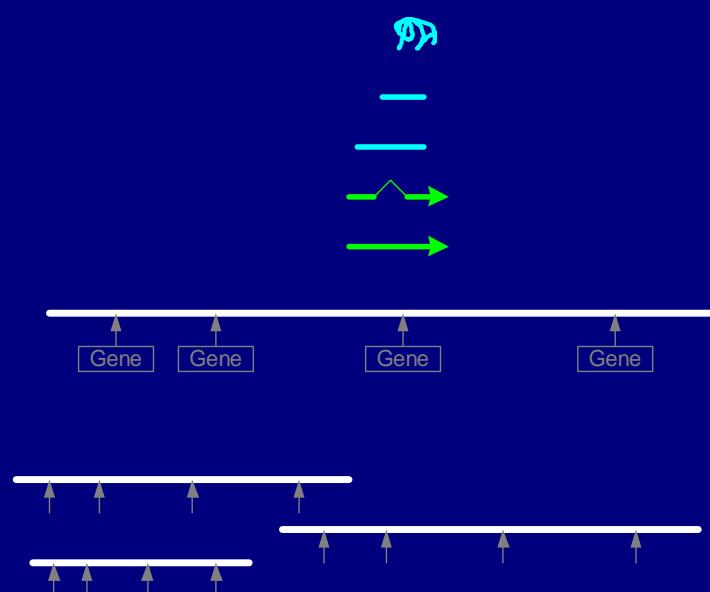
# The Basic Model



NCBI

WWW 110101

# Leveraging Resources





## The NCBI Data Model is defined in ASN.1

- ASN.1 is a data description language similar to a Backus-Naur Form.
- It is a formal language specifically designed to specify complex data structures in a machine, DBMS, and programming language independent manner.
- It is an international standard (ISO 8824, 8825)
- It is used by many data exchange protocols (e.g. X.400, Z39.50, WAIS).

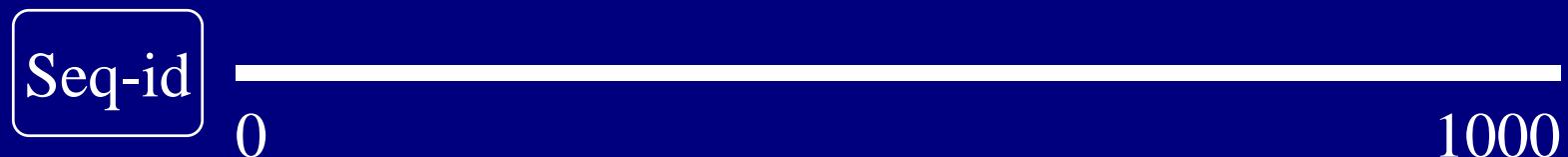


# A Bioseq defines an integer coordinate system.

## ■ ASN.1 definition

```
Bioseq ::= SEQUENCE {
    id            SET OF Seq-id,
    descr         Seq-descr           OPTIONAL,
    inst          Seq-inst ,
    annot         SET OF Seq-annot    OPTIONAL}
```

- The minimum required elements are an ID and the instance (e.g. length, topology, residues).



# There are many classes of Bioseq

- A Bioseq may be DNA, RNA, or protein.
- A Bioseq may be represented many ways.

virtual	.....	No residues
raw	-----	AGCCTTT
seg	-----.....-----	Parts by pointer
map	....↑.....↑.....↑..	Landmarks

- A Bioseq may have a history (Seq-hist)





## Seq-id's have different forms and usage

- Seq-id is defined as a choice of types with different forms and semantics.
- Some reflect the form and practice of the source databases or individuals.
- The NCBI “gi” is an arbitrary integer id which:
  - explicitly identifies a specific sequence
  - is stable and retrievable over time
  - has the same form over all sequence databases
  - is used to provide a history of changes to the sequence

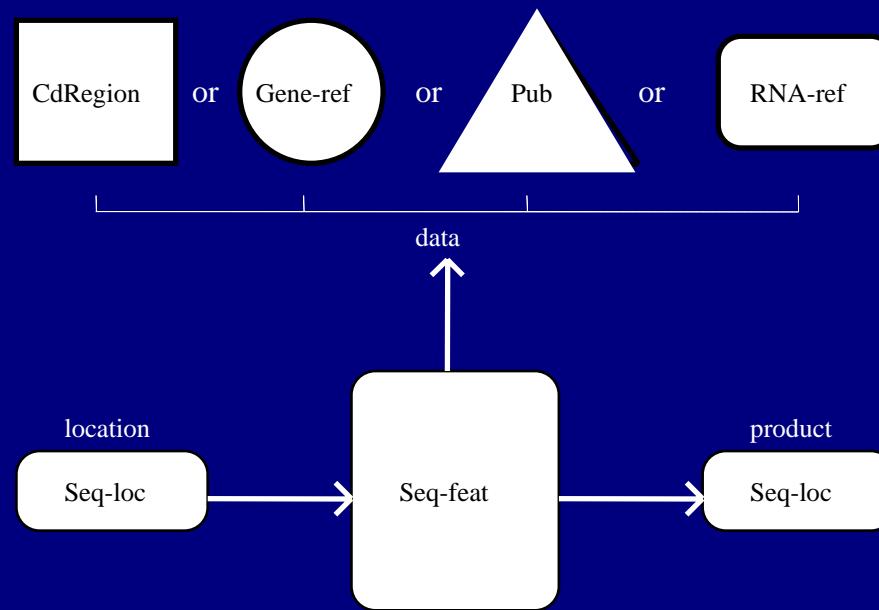


## A Seq-loc is a location on a Bioseq

- A Seq-loc is basically a Seq-id and an offset(s) on the coordinate system defined by that Seq-id
- Seq-loc is the same whether the Bioseq is DNA, protein; virtual, raw, segmented, or a map.
- A Seq-loc is explicit and not defined by context, meaning a Bioseq is easily referred to by data structures outside the entry.
- A Seq-align is simply a correlated set of Seq-loc.
- A Seq-align can be between any classes of Bioseq, sequence/sequence, sequence/map, etc.



# A Seq-feat links a Bioseq to non-sequence data and other databases



```
Gene-ref ::= SEQUENCE {  
    locus VisibleString OPTIONAL ,      -- Official gene symbol  
    allele VisibleString OPTIONAL ,     -- Official allele designation  
    desc VisibleString OPTIONAL ,       -- descriptive name  
    maploc VisibleString OPTIONAL ,     -- descriptive map location  
    pseudo BOOLEAN DEFAULT FALSE ,      -- pseudogene  
    db SET OF Dbtag OPTIONAL ,         -- ids in other dbases  
    syn SET OF VisibleString OPTIONAL } -- synonyms for locus
```

NCBI



EMBL

DDBJ

LANL

PDB

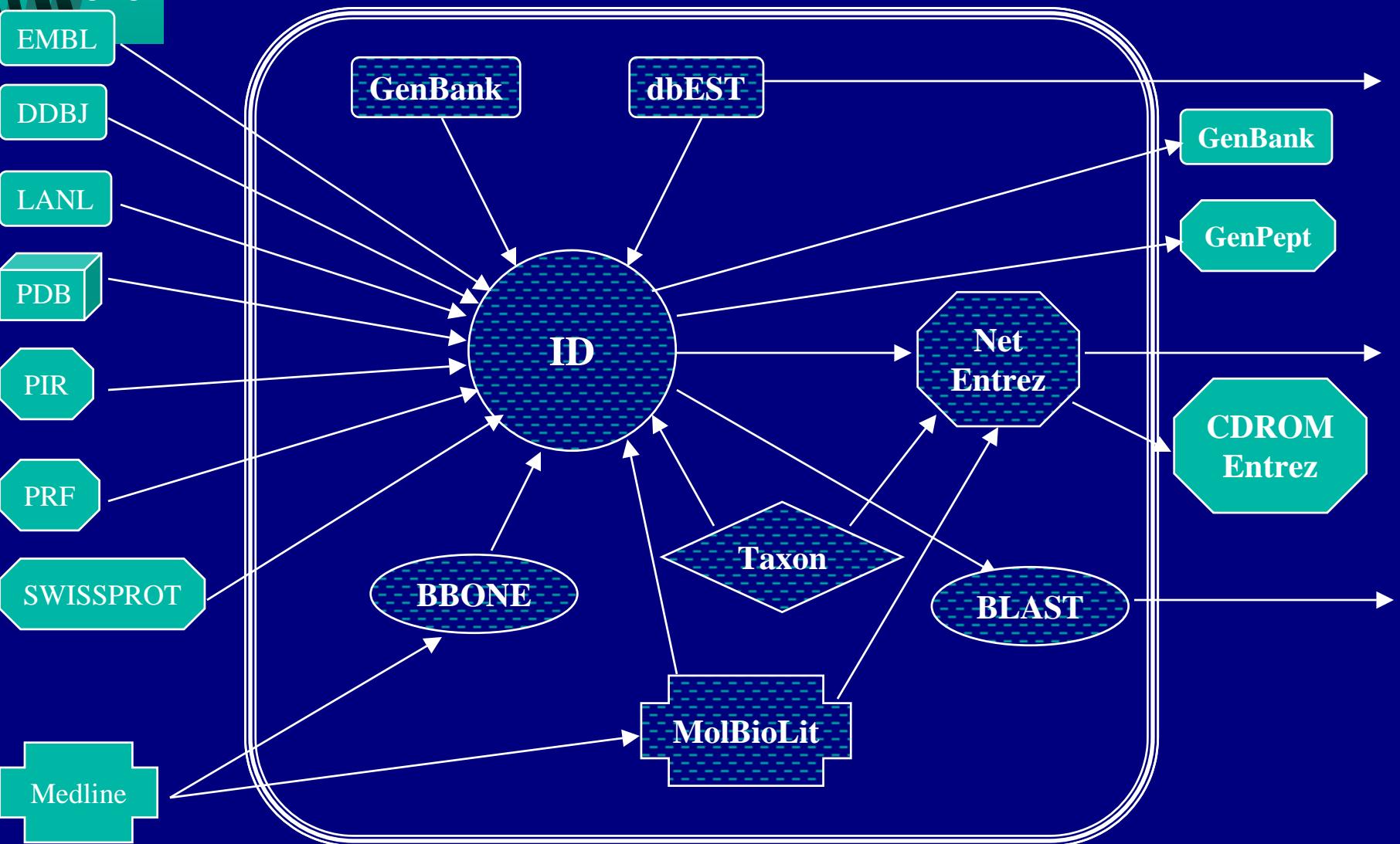
PIR

PRF

SWISSPROT

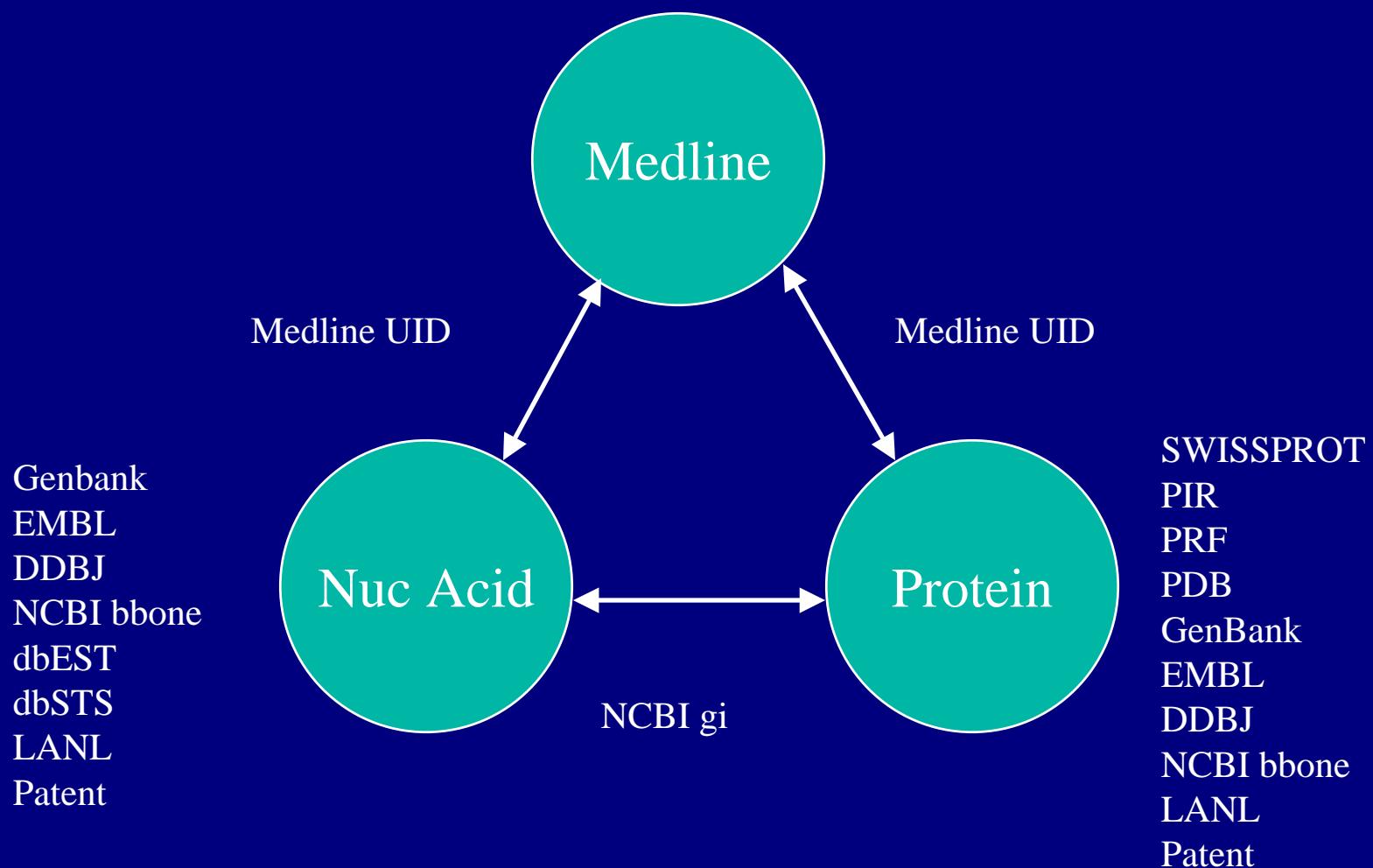
Medline

# Data Flow Through NCBI



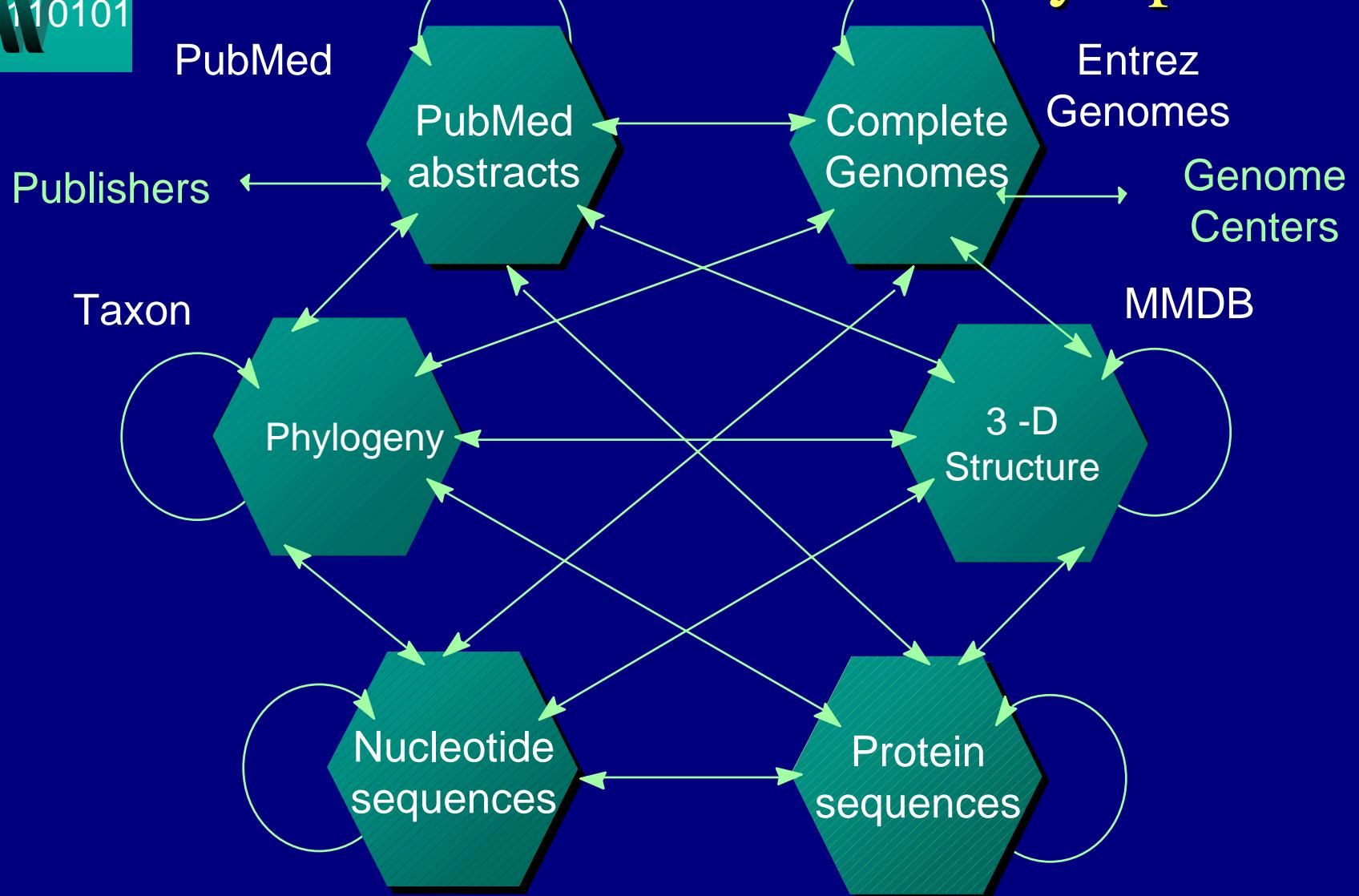


# Entrez integrates many different databases





# Entrez Increases Discovery Space



NCBI

WWW  
110101

# Sequence Variation and Structure

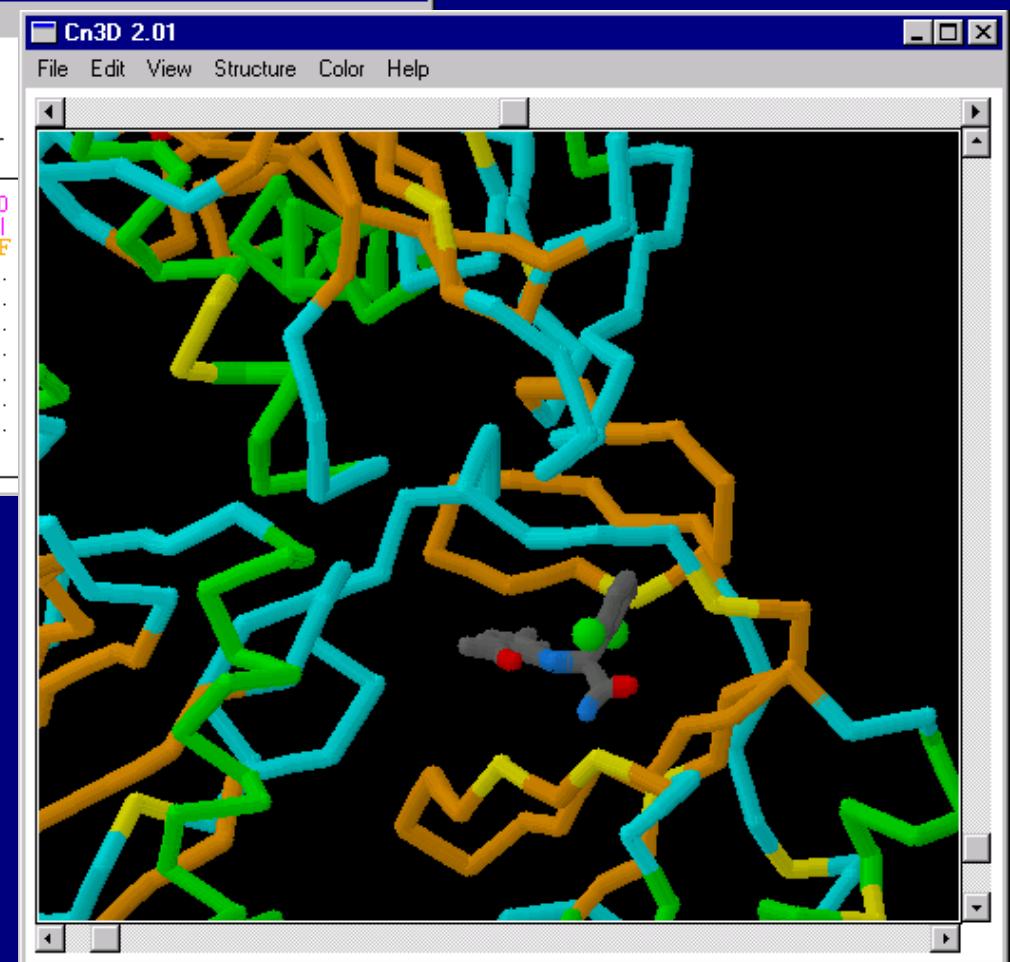
1VRU\_A

File Edit View Features Alignment

Go to: 0 Look at: 0

1VRU\_A Selections 89 - 89, 103 - 103, 108 - 108, 122 -

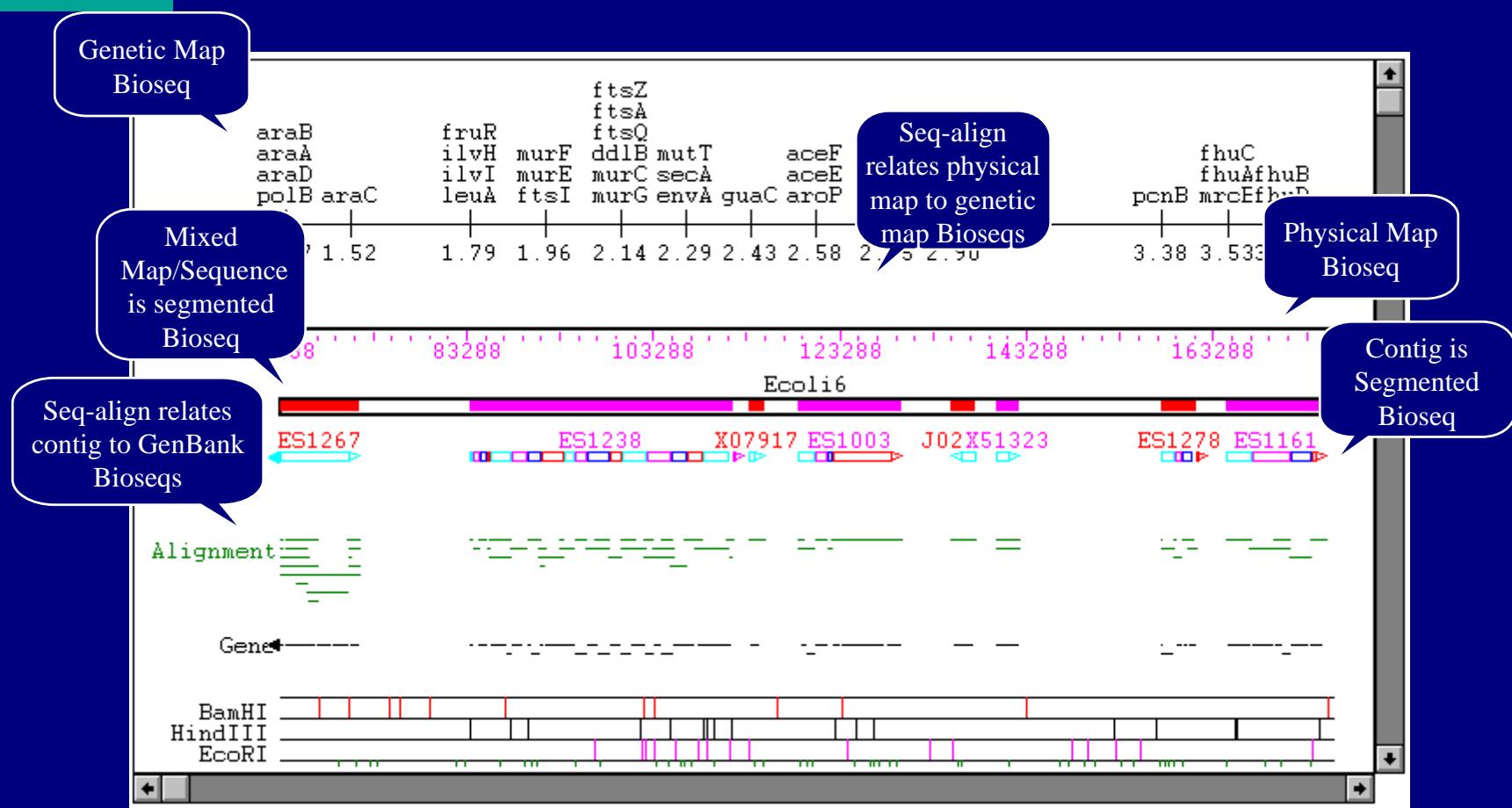
		110	120	130
1VRU_A	101	KKKKSVT	WLD	VGDAYFSVPL
Pyridin_R1	101	..N.....	.....	.....
Pyridin_R2	86	..N.....	.....	.....
Nevirap_R1	45	.....	.....	K
Nevirap_R2	45	.....	.....	K
MKC442_R1	45	.....I	.....	K
MKC442_R2	45	..R.....	.....	K
FoscarnetR	101	.....	.....	S



NCBI

110101

# Bioseqs and Seq-aligns build a comprehensive view on many levels



NCBI

WWW  
110101

# Fanconi Syndrome Genetic Mapping

Entrez-PubMed - Microsoft Internet Explorer

File Edit View Go Favorites Help

NCBI PubMed Nucleotide Protein Genome S

Search PubMed for fanconi syndrome genetic mapping

About Entrez

Entrez PubMed Overview Help | FAQ Tutorial New!Noteworthy

PubMed Services Journal Browser MeSH Browser Single Citation Matcher Batch Citation Matcher Clinical Queries Cubby

Related Resources Order Documents Grateful Med Consumer Health Clinical Alerts ClinicalTrials.gov PubMed Central

Privacy Policy

Entrez-PubMed - Microsoft Internet Explorer

File Edit View Go Favorites Help

NCBI PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search PubMed for fanconi syndrome genetic mapping

Limits Preview/Index History Clipboard

Display Abstract Save Text Order Add to Clipboard

1: Am J Hum Genet 2001 Jan;68(1):264-8 Related Articles, Books, OMIM, LinkOut  
**The University of Chicago Press**  
**Genetic and physical mapping of the locus for autosomal dominant renal Fanconi syndrome, on chromosome 15q15.3.**

**Lichter-Konecki U, Broman KW, Blau EB, Konecki DS.**  
Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, WI, USA.

Autosomal dominant renal Fanconi syndrome is a genetic model for the study of proximal renal tubular transport pathology. We were able to map the locus for this disease to human chromosome 15q15.3 by genotyping a central Wisconsin pedigree with 10 affected individuals. After a whole-genome scan with highly polymorphic simple sequence repeat markers, a maximum LOD score of 3.01 was calculated for marker D15S659 on chromosome 15q15.3. Linkage and haplotype analysis for an additional 24 markers flanking D15S659 narrowed the interval to approximately 3 cM, with the two highest single-point LOD scores observed being 4.44 and 4.68 (for D15S182 and D15S537, respectively). Subsequently, a complete bacterial artificial chromosome contig was constructed, from the High Throughput Genomic Sequence Database, for the region bounded by D15S182 and D15S143. The identification of the gene and gene product altered in autosomal dominant renal Fanconi syndrome will allow the study of the physiology of proximal renal tubular transport.

1:Lichter-Konecki U, Broman KW, Blau EB  
Genetic and physical mapping of the locus for autosomal dominant renal Fanconi syndrome, on chromosome 15q15.3.  
Am J Hum Genet. 2001 Jan;68(1):264-8.  
PMID: 11090339 [PubMed - indexed for MEDLINE]

2:Tay AH, Ren EC, Murugasu B, Sim SK, T  
Membranous nephropathy with an associated genetic defect.  
Pediatr Nephrol. 2000 Aug;14(8-9):747-51.  
PMID: 10955919 [PubMed - indexed for MEDLINE]

3:Nussbaum RL, Orison BM, Janne PA, C  
Physical mapping and genomic structure of the Fanconi anemia locus.  
Hum Genet. 1997 Feb;99(2):145-50.  
PMID: 9048911 [PubMed - indexed for MEDLINE]

4:Jean G, Fuchshuber A, Town MM, Gribois M, Hoff W, Niaudet P, Antignac C.  
High-resolution mapping of the gene for Fanconi anemia.  
Am J Hum Genet. 1996 Mar;58(3):535-43.  
PMID: 8644713 [PubMed - indexed for MEDLINE]

**NCBI**

110101

**Entrez Genome view - Microsoft Internet Explorer**

File Edit View Go Favorites Help

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy

Search for D15S182 OR D15S537 on chromosome(s) Find

Show linked entries Help FTP

**Homo sapiens genome view build 24**

Hits: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Hits: 14 15 16 17 18 19 20 21 22 X Y MT

**Entrez Map View - Microsoft Internet Explorer**

File Edit View Go Favorites Help

NCBI PubMed Entrez BLAST OMIM Taxonomy Structure

Search Advanced Search Map viewer Help Human Maps Help FTP Chr. 15 Resource Data As Table View Region Shown: 33325923 34325923 Go

**Homo sapiens Map View build 24**

**Chromosome:** 1 2 3 4 5 6 7 8 9 10 11 12 13 14 [ 15 ] 16 17 18 19 20 21 22 X Y

**Query:** D15S182 OR D15S537 [clear]

**Master: STS Map** Display settings

Total STSs On Chromosome: 3089 [23 not localized]  
Region Displayed: 33,325K-34,326K bp Download/View Sequence  
STSs Labeled: 20 Total STSs in Region: 31

marker Kbp

marker	Kbp
AFM261X5	33502
D15S516	33502
stSG51836	33558
D15S799	33572
mp0554	33572
sts-S82300	33575
GDB-454836	33575
WI-11756	33623
SHGC-11064	33668
SGC33828	33716
stSG4054	33726
SHGC-13414	33727
SHGC-30973	33755
D15S537	33766
sts-AA018839	33797
D15S182	33845
RH27235	33874
SHGC-6160	33993
SHGC-79674	34083
SHGC-5901	34228

Genes\_seq + Marshfield + STS marker

out zoom in

STS

SORD

B2M

LOC90538

FLJ21439

EIF3S1

HSPC129

symb: HSPC129  
orient.: -  
links: sv  
ev  
cyto.: 15q11.2  
full name: hypothetical protein

AFM261X5

D15S516

stSG51836

D15S799

mp0554

sts-S82300

GDB-454836

WI-11756

SHGC-11064

SGC33828

stSG4054

SHGC-13414

SHGC-30973

D15S537

sts-AA018839

D15S182

RH27235

SHGC-6160

SHGC-79674

SHGC-5901

AFM261X5

D15S516

stSG51836

D15S799

mp0554

sts-S82300

GDB-454836

WI-11756

SHGC-11064

SGC33828

stSG4054

SHGC-13414

SHGC-30973

D15S537

sts-AA018839

D15S182

RH27235

SHGC-6160

SHGC-79674

SHGC-5901

AFM261X5

D15S516

stSG51836

D15S799

mp0554

sts-S82300

GDB-454836

WI-11756

SHGC-11064

SGC33828

stSG4054

SHGC-13414

SHGC-30973

D15S537

sts-AA018839

D15S182

RH27235

SHGC-6160

SHGC-79674

SHGC-5901

**NCBI**

**Blast Result - Microsoft Internet Explorer**

File Edit View Go Favorites Help

**NCBI Sequence Viewer - Microsoft Internet Explorer**

gi|7705461 view - Microsoft Internet Explorer

File Edit View Go Favorites Help

**Blast 2 Sequences results**

BLAST PubMed Nucleotide Protein Genome Structure Taxonomy Help

Query: gi|7705461 hypothetical protein [Homo sapiens]  
 Matching gi: 6841480  
 Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo

Best hits Common Tree Taxonomy Report 3D structures CDD-Search GI list

116 BLAST hits to 14 unique species Sort by taxonomy proximity

0 Archaea 0 Bacteria 62 Metazoa 27 Fungi 21 Plants 1 Viruses 5 Other Eukaryotae

Keep only ▾ Cut-Off 100 Select Reset

466 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
2378	26	XP_032383	14785185	hypothetical protein [Homo sapiens]
1825	26	AAF29030	6841354	HSPC058 [Homo sapiens]
1490	26	XP_007648	14785187	14635 [Homo sapiens]
1489	26	BAA91664	7022613	unnamed protein product [Homo sapiens]
737	7	AAF60646	7331958	contains similarity to several yeast and human hypothetical proteins [Caenorhabditis elegans]
630	3	CAB87659	7573353	putative protein [Arabidopsis thaliana]
630	3	T48545	11282308	hypothetical protein F14F18.30 - Arabidopsis thaliana
449	3	BAB19125	11761135	putative HSPC058 [Oryza sativa]
406	18	AAF17484	6572958	NLI-interacting factor isoform R5; NLI/Ldb1/CLIM interacting factor [Gallus gallus]
406	18	AAF17482	6572954	NLI-interacting factor isoform T2; NLI/Ldb1/CLIM interacting factor [Gallus gallus]
405	26	BAA21667	2289786	HYA22 [Homo sapiens]
405	26	NP_005799	5031775	HYA22 protein [Homo sapiens]
403	3	AAD28548	4731912	development protein DG1148 [Dictyostelium discoideum]
399	4	CAA97541	1360322	ORF YLR019w [Saccharomyces cerevisiae]
399	4	S64841	2131751	hypothetical protein YLR019w - yeast (Saccharomyces cerevisiae)
399	4	NP_013119	6323047	Per2p [Saccharomyces cerevisiae]
392	18	AAF17481	6572952	NLI-interacting factor isoform T1; NLI/Ldb1/CLIM interacting factor [Gallus gallus]
388	26	JC5707	7512494	HYA22 protein - human
388	4	CAA97454	1360175	ORF YLL010c [Saccharomyces cerevisiae]
388	4	NP_013091	6323019	Psrip [Saccharomyces cerevisiae]
388	4	CAA62782	1495214	L1341 protein [Saccharomyces cerevisiae]
388	4	S64752	2131724	hypothetical protein YLL010c - yeast (Saccharomyces cerevisiae)
384	21	AAK83555	15145799	golli-interacting protein [Mus musculus]
384	26	AAG15402	10257407	nuclear LIM interactor-interacting factor [Homo sapiens]
384	26	AAG15404	10257410	nuclear LIM interactor-interacting factor [Homo sapiens]

Internet

NCBI

WWW 110101

Entrez-PubMed - Microsoft Internet Explorer

File Edit View Go Favorites Help

Entrez-PubMed - Microsoft Internet Explorer

File Edit View Go Favorites Help

NCBI PubMed National Library of Medicine NLM

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard

Display Abstract Save Text Order Add to Clipboard

□ 1: J Biol Chem 2000 Jun 23;275(25):19352-60 Related Articles, Books, LinkOut  
FREE full text article at [www.jbc.org](http://www.jbc.org)

**Psr1p/Psr2p, two plasma membrane phosphatases with an essential DXDX(T/V) motif required for sodium stress response in yeast.**

Siniosoglou S, Hurt EC, Pelham HR.

Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom.

Regulation of intracellular ion concentration is an essential function of all cells. In this study, we report the identification of two previously uncharacterized genes, PSR1 and PSR2, that perform an essential function under conditions of sodium ion stress in the yeast *Saccharomyces cerevisiae*. Psr1p and Psr2p are highly homologous and were identified through their homology with the endoplasmic reticulum membrane protein Nem1p. Localization and biochemical fractionation studies show that Psr1p is associated with the plasma membrane via a short amino-terminal sequence also present in Psr2p. Growth of the psr1psr2 mutant is severely inhibited under conditions of sodium but not potassium ion or sorbitol stress. This growth defect is due to the inability of the psr1psr2 mutant to properly induce transcription of ENA1/PMR2, the major sodium extrusion pump of yeast cells. We provide genetic evidence that this regulation is independent of the phosphatase calcineurin, previously implicated in the sodium stress response in yeast. We show that Psr1p contains a DXDX(T/V) phosphatase motif essential for its function in vivo and that a Psr1p-PtA fusion purified from yeast extracts exhibits phosphatase activity. Based on these data, we suggest that Psr1p/Psr2p, members of an emerging class of eukaryotic phosphatases, are novel regulators of salt stress response in yeast.



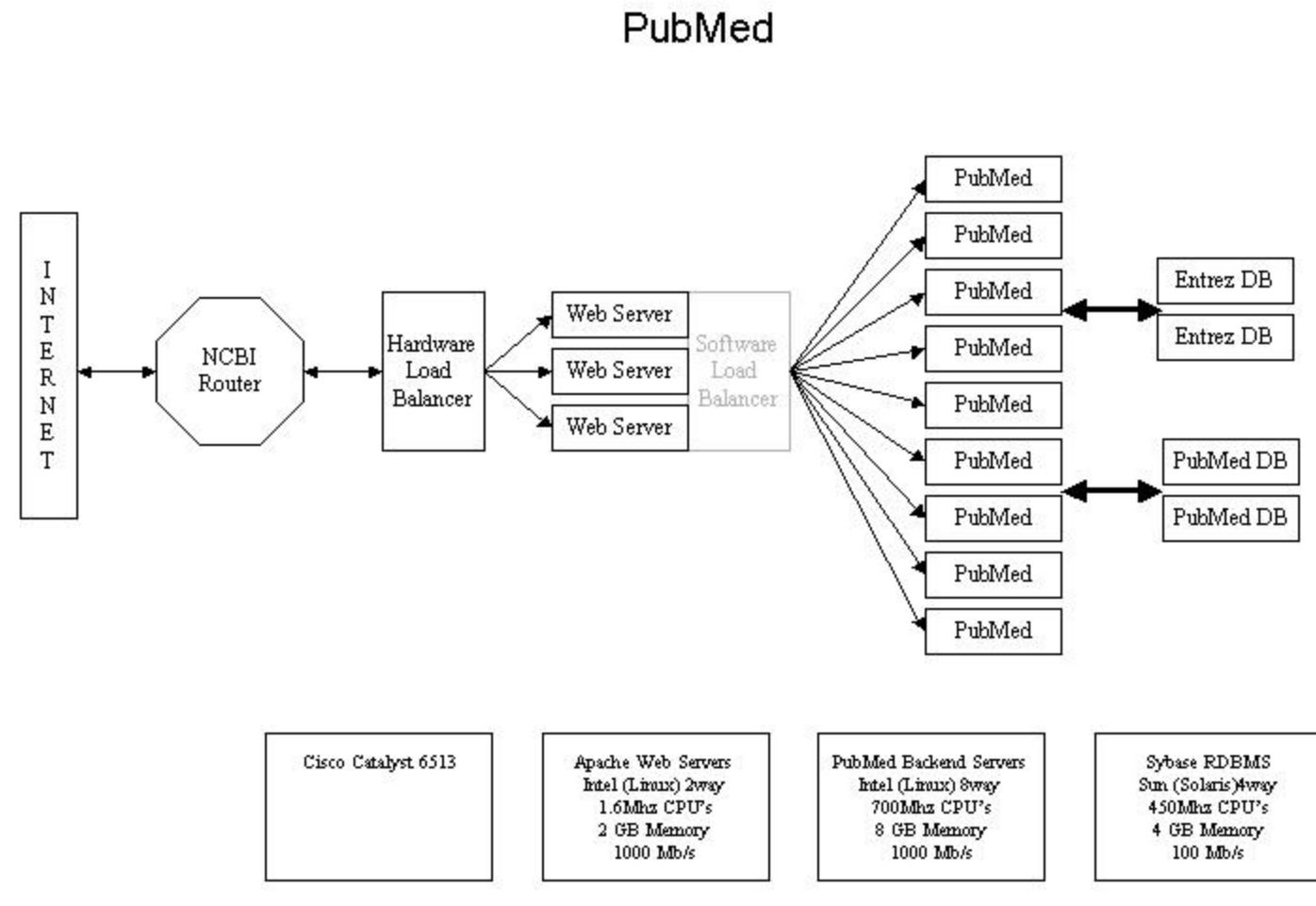
# NLM Annotation Project

PubMed articles with comment attached to gene products as part of the MeSH indexing flow.

In first 18 months, almost 2000 articles a month annotated on 8 organisms.

	Genes	Articles
Human	5981	20945
Mouse	3864	9675
Rat	1850	4256
Fly	824	1020
Nematode	134	118
Zebrafish	189	235
Cow	176	201
Total	13027	34549

# PubMed Hardware





# CPU Growth on 3 Services

## ■ NCBI Intel Compute Farm:

■ January 2000	6 – Intel (Solaris) 8ways, 550 MHz, 8GB	48 CPUs	26,400 MHz
■ January 2002	19 – Intel (Solaris) 8ways, 550 MHz, 8GB	152 CPUs	83,600 MHz
■ June 2003	84 – Intel (Linux) 2ways, 1.6 GHz, 4GB	168 CPUs	268,800 MHz

## ■ BLAST Computer Farm:

■ January 2000	8 – Intel (Solaris) 4ways, 400MHz, 4GB	32 CPUs	12,800 MHz
■	5 – Intel (Solaris) 8ways, 700MHz, 8GB	40 CPUs	28,000 MHz
■	3 – SGI (IRIX) 12 ways, 250 MHz, 2GB	36 CPUs	9,000 MHz 49,800 MHz
■ January 2002	22 – Intel (Solaris) 8ways, 700 MHz, 8GB	176 CPUs	123,200 MHz 123,200 MHz
■ June 2003	22 – Intel (Linux) 8ways, 700 MHz, 8GB 50 – Intel (Linux) 2ways, 1.6 GHz, 4GB	176 CPUs 100 CPUs	123,200 MHz 160,000 MHz 283,200 MHz

## ■ PubMed Backend Servers

■ January 2000	6 – Sun Enterprise Servers, 400 MHz, 4GB	48 CPUs	19,200 MHz
■ January 2002	6 – Intel (Linux) 8ways, 700 MHz, 8GB	48 CPUs	33,600 MHz
■ June 2003	12 – Intel (Linux) 8ways, 700 MHz, 8GB	96 CPUs	67,200 MHz



## Stable Identifiers

- Accession.version, GI number
- PubMed ID, PMC ID
- TaxId
- LocusLink ID, Geneid
- CDD id
- SNP id, RSNP id
- OMIM number



# Coding Standards

- ANSI C
  - C Toolkit
- ANSI C++
  - C++ Toolkit
- HTTP
  - HTML
  - Web Services
- XML
  - XSLT
  - XML Schema
- SQL



# Data Exchange

- GenBank Flatfile
  - GBSeq XML
  - Fasta
- AGP, GFF, 5 column feature table
- ASN.1
- Paup, Phylip, etc
- MEDLINE
  - NLM DTD
- PMC
  - Publishing and Archiving DTD



# Bioinformatics Survival

- A guiding vision helps choose the best path
- The journey of 1000 miles starts with a single step
  - or
- Just because something is doable is no reason not to do it
- Standards are good when they are yours
- The secret of happiness is to not expect too much