Informatics in Biological Research

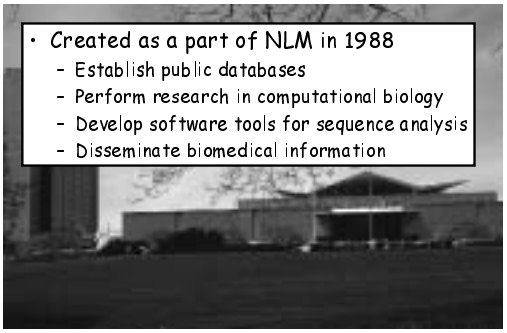
Agenda for Informatics in the Biological Sciences

University of Chicago *13 Jan 2003*

Peter Cooper
National Center for Biotechnology Information

NCBI

The National Center for Biotechnology Information



- Created as a part of NLM in 1988
  - Establish public databases
  - Perform research in computational biology
  - Develop software tools for sequence analysis
  - Disseminate biomedical information

NCBI

The Revolutions

- Biology
  - DNA sequencing
  - High throughput methodologies
  - Improvements in cloning techology
  - PCR
  - Microarray techniques
- Computer Science
  - CPU speed
  - Storage devices
  - The Internet and the WWW

NCBI

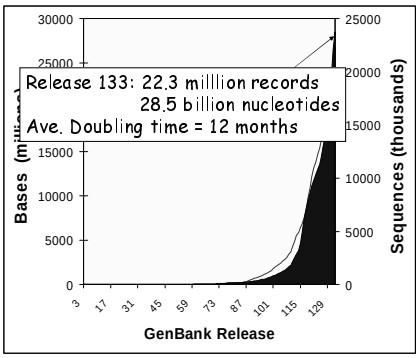
The Selection Pressure

"It's sink or swim as a tidal wave of data approaches"

Nature 399:517 10 June 1999

NCBI

The Growth of GenBank

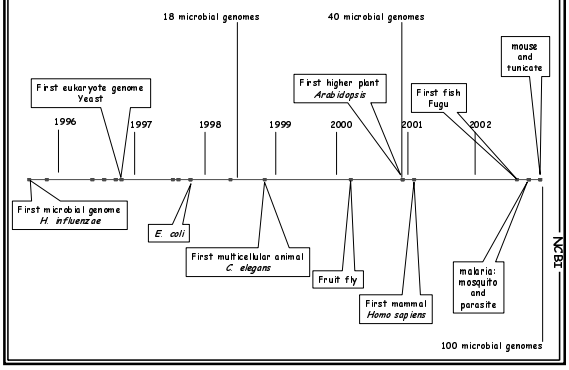


NCBI

The Genome Sequencing Era

NCBI

The Genome Sequencing Era



NCBI

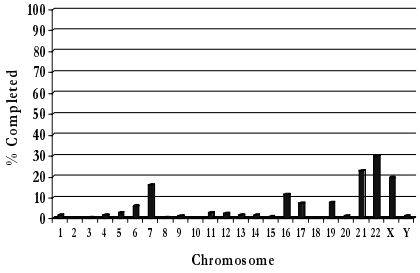
Coming soon …

- Nearly done
  - rat
  - purple sea urchin
  - zebrafish
- NHGRI's Priority Organisms
  - Chicken
  - Cow
  - Dog
  - Chimpanzee
  - Honeybee
  - *Tetrahymena*
  - *Oxytichia*
  - Several fungi
- Over 100 bacterial genomes

NCBI

Human Genome Sequencing Progress

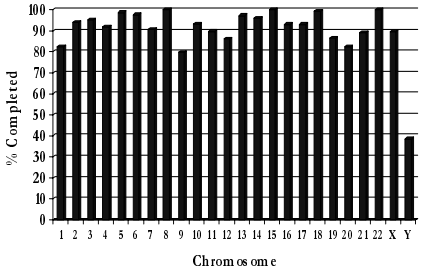
June 1998



NCBI

Human Genome Sequencing Progress

Dec. 2002



NCBI

The Rapid Evolution of Resources

NCBI

Web Access

Text

Entrez

Sequence

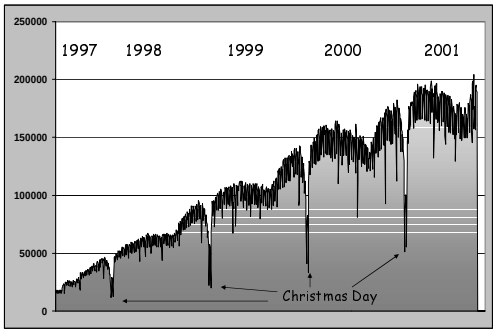
BLAST

Structure

VAST

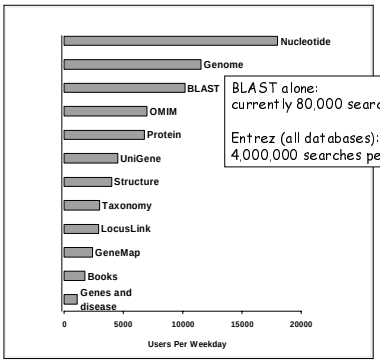
NCBI

Users per day



NCBI

Some Web Statistics

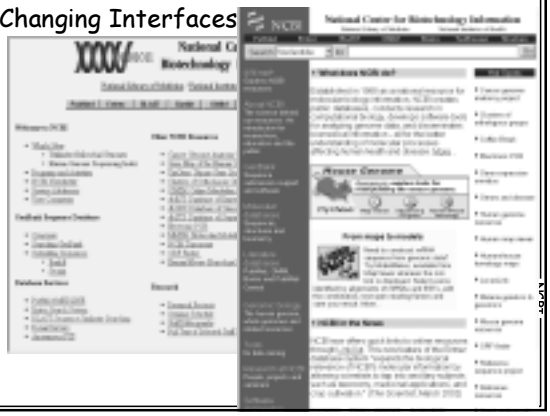


BLAST alone:
currently 80,000 searches per day

Entrez (all databases):
4,000,000 searches per day

NCBI

Changing Interfaces

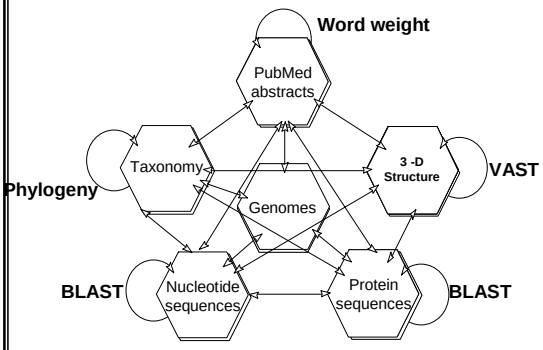


NCBI

Changing Interfaces

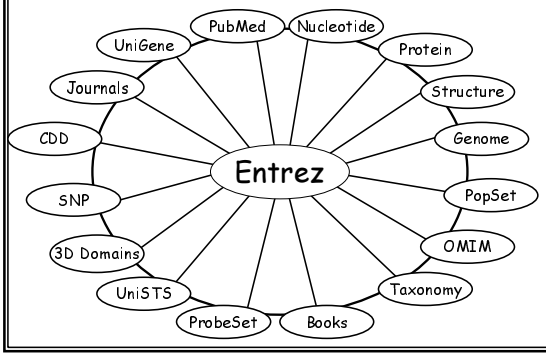
NCBI

Entrez: Database Integration



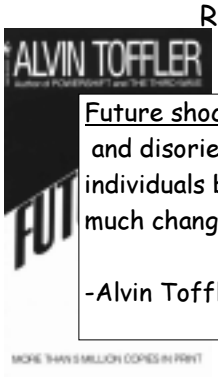
NCBI

The (ever) Expanding Entrez System



NCBI

Revolution



Future shock: the shattering stress and disorientation that we induce in individuals by subjecting them to too much change in too short a time.

-Alvin Toffler

NCBI

The Promise of Computational Biology

Gene

```
> DNA sequence
AATTCATGAAAATCGTATACTGGTCTGGTACCGGCAACAC
TGAGAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA
TCTGGTAAAGACGTCAACACCATCAACGTGTCTGACGTTA
ACATCGATGAACTGCTGAACGAAGATATCCTGATCCTGGG
TTGCTCTGCCATGGGCGATGAAGTTCTCGAGGAAAGCGAA
TTTGAACCGTTCATCGAAGAGATCTCTACCAAAATCTCTG
GTAAGAAGGTTGCGCTGTTCGGTTCTTACGGTTGGGGCGA
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC
TACGGTTGCGTTGTTGTTGAGACCCCGCTGATCGTTCAGA
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAATTTGG
TAAGAAGATCGCGAACATCTAGTAGA
```

Structure & Function

```
> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI
```

The power of computing on the data

NCBI

The Cross-over to "Functional Genomics"

"In the past we have had functions in search of sequences. In the future, pathology and physiology will become 'functionators' *for the sequences*."

Daniel C. Tosteson, Dean
Harvard Medical School
March 26, 1997

The future is NOW.

NCBI

Comparative Analysis of Genomes

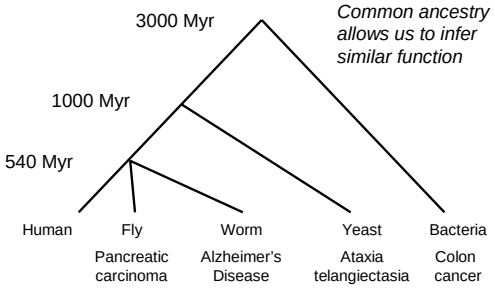
NCBI

"What is true for *E. coli* is also true for elephant." Jacques Monod, c. 1961

"What is true for yeast is also true for human." David Botstein, 1988

NCBI

Molecular Evolution



NCBI

"Homology...

... is the central concept for *all* of biology. Whenever we say that a mammalian hormone is the 'same' hormone as a fish hormone, that a human gene sequence is the 'same' as a sequence in a chimp or a mouse, that a HOX gene is the 'same' in a mouse, a fruit fly, a frog, and a human -- even when we argue that discoveries about a worm, a fruit fly, a frog, a mouse, or a chimp have relevance to the human condition -- we have made a bold and direct statement about homology. The aggressive confidence of modern biomedical science implies that we know what we are talking about."

David B. Wake

NCBI

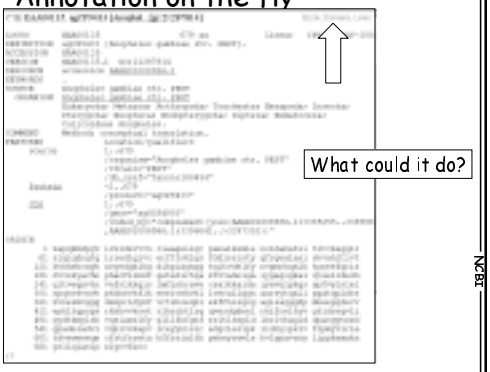
Application:
MLH1 Homolog in *Anopheles gambiae*

NCBI

Annotation on the fly: mosquito genome

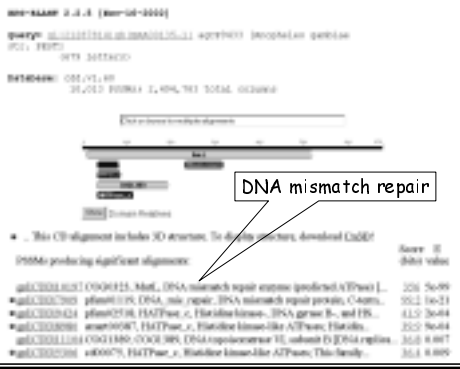
NCBI

Annotation on the fly



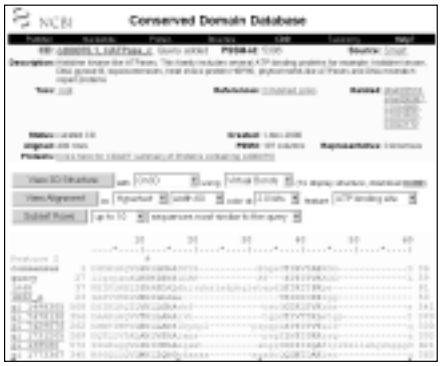
NCBI

Conserved Domains



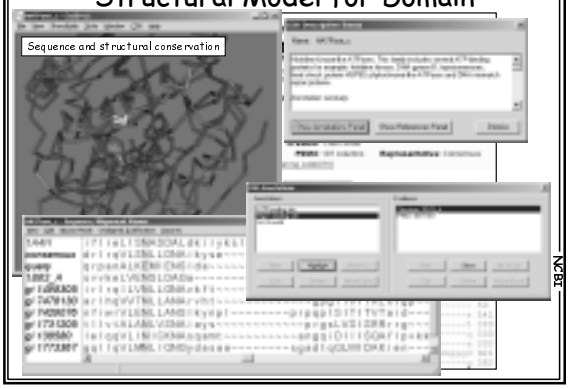
NCBI

NCBI Curated CD



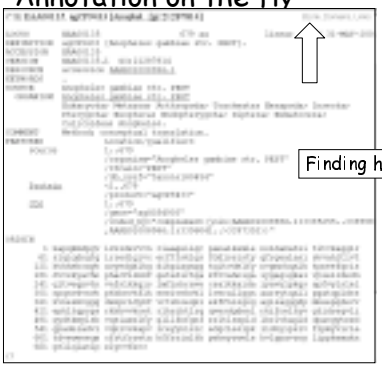
NCBI

Structural Model for Domain



NCBI

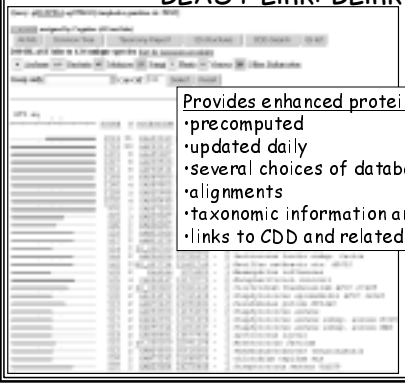
Annotation on the fly



Finding homologues

NCBI

BLAST Link: BLink



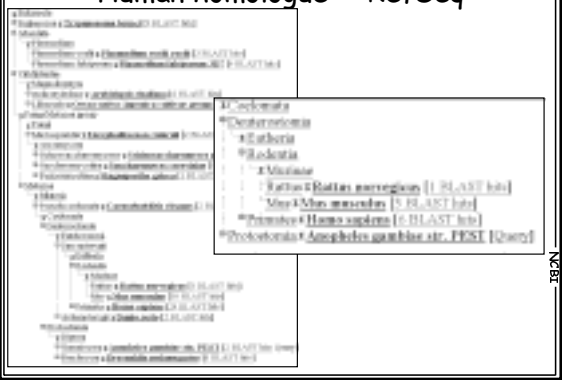
Provides enhanced protein BLAST output
- precomputed
- updated daily
- several choices of database
- alignments
- taxonomic information and sorting
- links to CDD and related structures

NCBI

BLAST Link: BLink

NCBI

Human homologue -> RefSeq



NCBI

NCBI Curated Proteins

NCBI

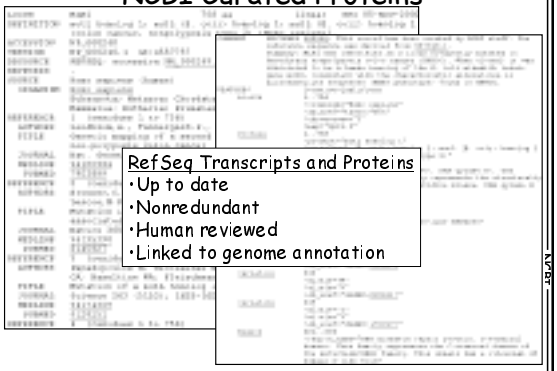
NCBI Curated Proteins

NCBI

NCBI Curated Proteins

NCBI

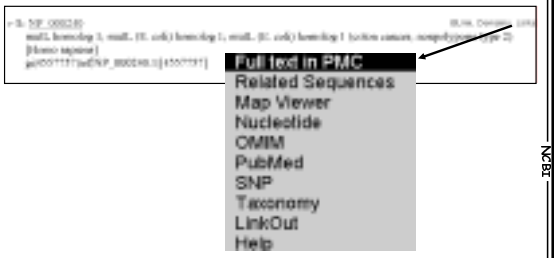
NCBI Curated Proteins



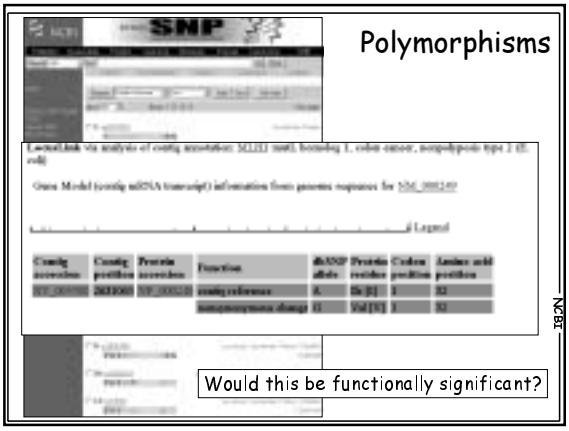
RefSeq Transcripts and Proteins
- Up to date
- Nonredundant
- Human reviewed
- Linked to genome annotation

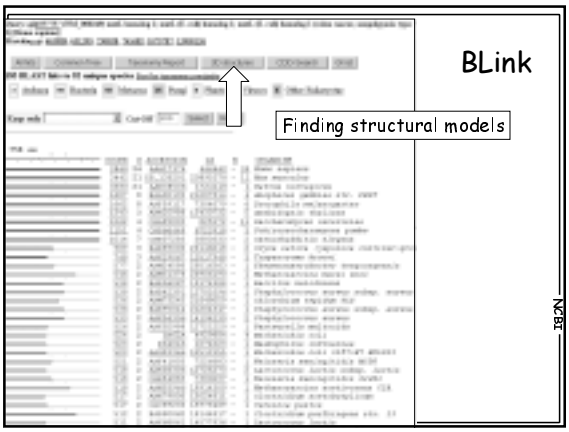
NCBI

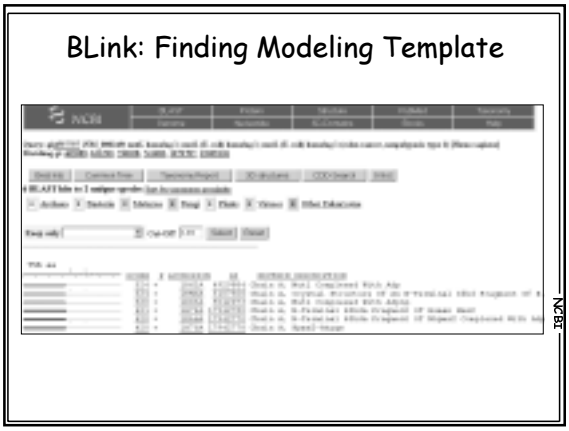
Links and Neighbors Human mutL

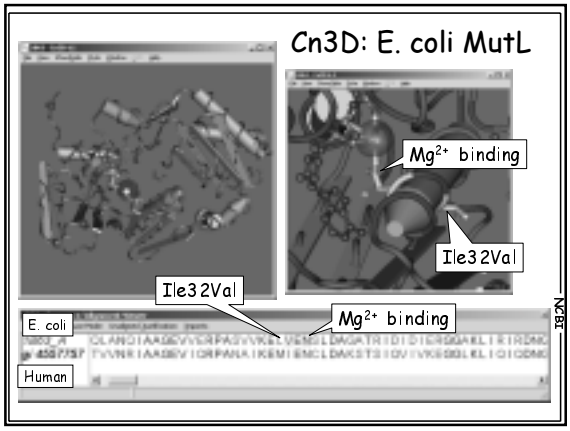


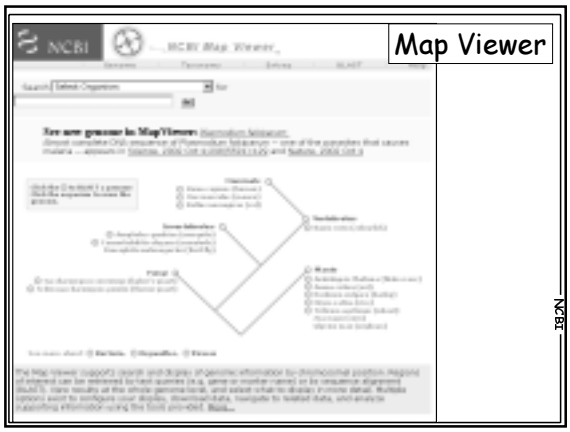
NCBI

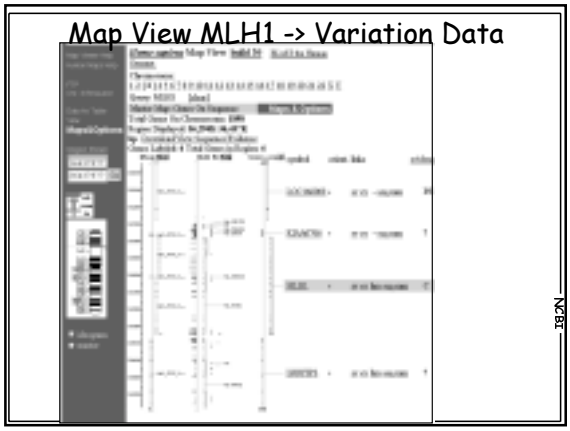

Polymorphisms
Would this be functionally significant?
NCBI


BLink
Finding structural models
NCBI

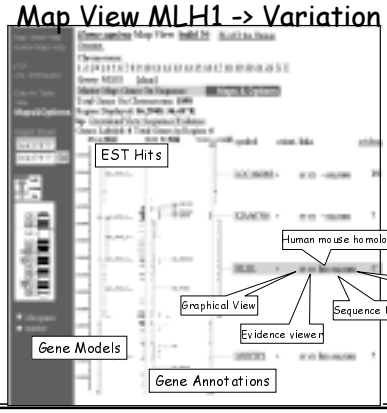

BLink: Finding Modeling Template
NCBI


Cn3D: E. coli MutL
Mg2+ binding
Ile32Val
Ile32Val
Mg2+ binding
E. coli
Human
NCBI


NCBI
Map Viewer
NCBI

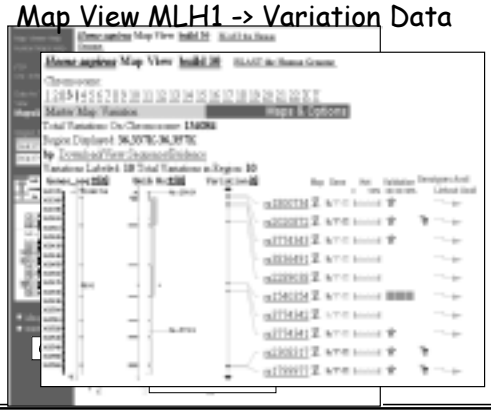

Map View MLH1 -> Variation Data
NCBI

Map View MLH1 -> Variation Data



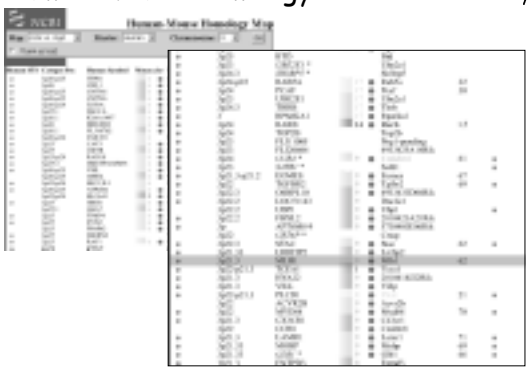
NCBI

Map View MLH1 -> Variation Data



NCBI

Human Mouse Homology: the mouse assembly



NCBI

The Mouse Assembly

NCBI

NCBI Training

NCBI

Field Guide Content

One day practical training course

- Molecular Databases
  - GenBank
  - RefSeq
- Web Access
  - Entrez
  - BLAST
- Genomic Biology

NCBI

Summary Aug 1998 - Dec 2002

- 90 Field Guides
- 3 BLAST courses
- 17 short courses
- 36 States plus Puerto Rico
- 6000 participants

NCBI

Field Guide Course

NCBI

Service Addresses

•*General Help*	`info@ncbi.nlm.nih.gov`
•*BLAST*	`blast-help@ncbi.nlm.nih.gov`

NCBI