

Biological Databases and Tools

Sandra Sinisi / Kathryn Steiger
November 25, 2002

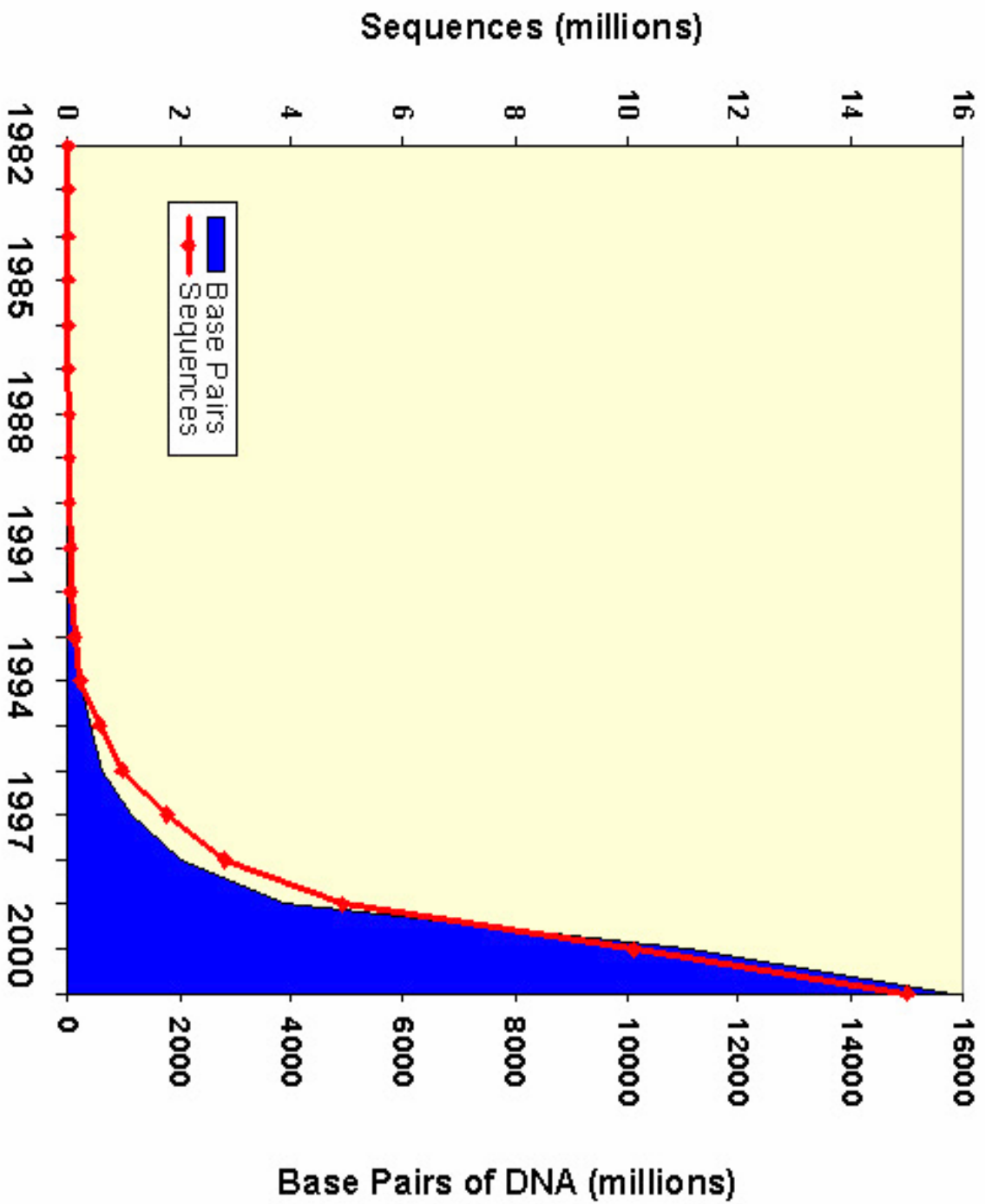
Introduction

- More than storage
- Qualities of a good database
 - ◆ Flexible retrieval
 - ◆ Analysis software compatible
 - ◆ Data cleaning features

The need for electronic access

- Quantity of data has grown
- Data concentrated in distant locales
- Field is quickly developing so we need to relate new information to existing data

Growth of GenBank



Types of Data

- Nucleotide sequences
- Protein sequences
- Protein structure
- Functional
- Secondary source information

Public Nucleotide Sequence Sites

- EMBL European Molecular Biology Laboratory nucleotide database from the European Bioinformatics Institute (EBI, Hinxton, UK)
 - ◆ <http://www.ebi.ac.uk/embl/>
- The Institute for Genomic Research (Rockville, MD)
<http://www.tigr.org/tdb/>

- DDBJ (Mishima, Japan)

DNA Data Bank of Japan

- ◆ <http://www.nig.ac.jp/home.html>

NCBI, DDBJ, and EMBL provide separate points of data submission, yet exchange this information daily, making the same database (in different formats and information systems) available to the community at-large.



Protein Sequence Databases

- SwissProt Integrated with other databases
 - ◆ <http://www.ebi.ac.uk/swissprot/>
- TrEmbl Translation of nucleotide sequences into protein sequences
 - ◆ <http://www.expasy.org/sprot/sprot-top.html>

Protein 3D Structure

- PDB Protein Databank
 - ◆ <http://nist.resb.org/pdb/>
- BioMagResBank
 - ◆ <http://bimas/dcrt.nih.gov>
- Structural Classification of Proteins
 - ◆ <http://scop.mrc-lmb.cam.ac.uk/scop/>

A good starting point ...

- National Center for Biotechnology Information:
<http://www.ncbi.nlm.nih.gov>

SITE MAP

- About NCBI
- GenBank
- Molecular databases
- Literature databases
- Genomic biology
- Tools
- Research at NCBI
- Software engineering
- Education
- FTP site
- Contact information

Hot Spots




- ▶ Cancer genome anatomy project
- ▶ Clusters of orthologous groups
- ▶ Coffee Break
- ▶ Electronic PCR
- ▶ Gene expression omnibus
- ▶ Genes and disease
- ▶ Human genome resources
- ▶ Human map viewer
- ▶ Human/mouse homology maps
- ▶ LocusLink
- ▶ Malaria genetics & genomics
- ▶ Mouse genome resources
- ▶ ORF finder
- ▶ Reference sequence project
- ▶ Retrovirus resources
- ▶ Serial analysis of gene expression
- ▶ SKYCOH database

▶ What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

Mouse Genome

Resources: explore tools for manipulating the mouse genome.

Try these!  Map Viewer  Sequencing Progress  Human-Mouse Homology

BLink and get results fast!

Use BLink to view a graphical alignment of protein sequence similarities, taxonomic trees, 3D structures, and more. BLink provides quick results based on precomputed BLASTp searches against the non-redundant (nr) protein database. [More...](#)

New malaria mosquito resource

A new web page devoted to *Anopheles gambiae* is now available. Access both the original sequence data deposited in GenBank as well as the first version of the genome assembly. BLAST the *Anopheles gambiae* genome and browse links to information on taxonomy. [More...](#)

▶ NCBI News

October 2002 marks the 20th anniversary of the creation of [GenBank](#). GenBank has grown



National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

PubMed

Entrez

BLAST

OMIM

Books

TaxBrowser

Structure

Search for

- PubMed – gateway to biomedical research literature
- Entrez – search engine
- BLAST – most important
- OMIM – Online Mendelian Inheritance in Man database
- Taxonomy – groups all data by taxonomic classification
- Structure – contains the 3D structure for all nucleic acids & proteins whose shape has been determined by X-ray or NMR

Basic Local Alignment Search Tool (BLAST) program

- Important software tool for searching sequence databases
- Can be used to search databases using nucleic acid or protein query sequences
- Allows dynamic search of the sequence databases to find similar sequences in different organisms

Accepted input types

- FASTA format

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (" $>$ ") symbol in the first column.

- GenBank format

“DNA-centered” view of a sequence record.

Protein Sequence Databases: Format

Genbank format:

```
LOCUS Sc_16 7000bp DNA PLN:08.MAY.2000
DEFINITION Saccharomyces cerevisiae chromosome XVI strain S288C.
ACCESSION Sc_16
VERSION
KEYWORDS .
SOURCE baker's yeast
ORGANISM Saccharomyces cerevisiae Eukaryota
REFERENCE 1 (bases 1 to 7000)
AUTHORS Goffeau A, et al
TITLE Direct Submission
JOURNAL Submitted (08-MAY-2000) NCBI/NLM, National Institutes
of Health, Building 38A, Room 8N905, Bethesda, MD 20894, USA
FEATURES Location/Qualifiers
translation="DHNGTIVHKSQDVPPIHKIPKSLIHDDQDINPYNGSENE
RKPNIERKDYDRVGDPPMRMDRYGTYTLLKPKQKELTYQLKRFG....
"
BASE COUNT 2201 a 1276 c 1255 g 2268 t
ORIGIN
```

1 gcccaccant gggacgatt g tccatnnt c agggantgt cctnctcna tnaagatnc
6t aatccagatct ctanlcnatg accatgatat caatctctt aatggthccg aaatcgaag
12t aatccatnt ctatgacgta gaaagctcga ccggttgtt gtt ccantga gaaatgattg
[ctc.]

FASTA format:

```
>gls32319|gtdrTFV2E|TFV2E: envelope
protein
ELRLRYCAPAGFALLKCNDAVDGFKT
NCSNVSVHCTNLNNTVTYTGILLNGS
YSENRTOIWQKHRTSNDLALLLNKHY
NLTYTCKRPGNKTYLPTVMAGLVPHS
QKYNLRLRQAWCHEPSSWKGAWKEY
KEELVNLPKERYRG:TNDRKRLFQROW
GDPEFANLWFNCHGEFFYCKAIDWFLN
YLNMLTYDADHNECKNTSGTKSGNKR
APGPGVQRITYVACHIRS VIMWLETTSKK
TYAAPREGHLECTSTVTCMTVELNYIP
KNRTVYTLSPQESISWAELDRYKLYEI
TPIGFAPTEVRRYTGCHERQKRVPVYX
XXXXXXXXXXXXXXXXXXXXVQSQDH
LLAGHLQQQKKNLAAVEAQQQMLKLLI
WGVK
```

>ggl....

most software tools accept

FASTA or GenBank formats

BLAST Query Results of insulin protein from the Zebra fish

Your request has been successfully submitted and put into the Blast Queue.

Query = gi|12053668|emb|CAC20109.1| insulin [Danio rerio] (108 letters)

Putative conserved domains have been detected, click on the image below for detailed results.



Blast It!!

[Search](#)

```
>gi|1 2053668|emb|CAC20109.1|insulin[Danio  
rerio]MAVWIQAGALLVLLWSSVSTNPGTPQHLGSHLV  
DALYLVCGPTGFFYNPKRDVEPLLGLPCKSAQETVA  
DFAFKDHAELIRKRGIVEQCCHKPCSIFELQNYCN
```

[Set subsequence](#)

From: To:

[Choose database](#)

nr

[Do CD-Search](#)



Now:

BLAST!

or

Reset query

Reset all

- ✓ nr
- swissprot
- ✓ pat
- yeast
- ecoli
- pdb
- Drosophila genome
- month

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show define and scores. Click to show alignments

Color Key for Alignment Scores

<40

40-50

50-80

80-200

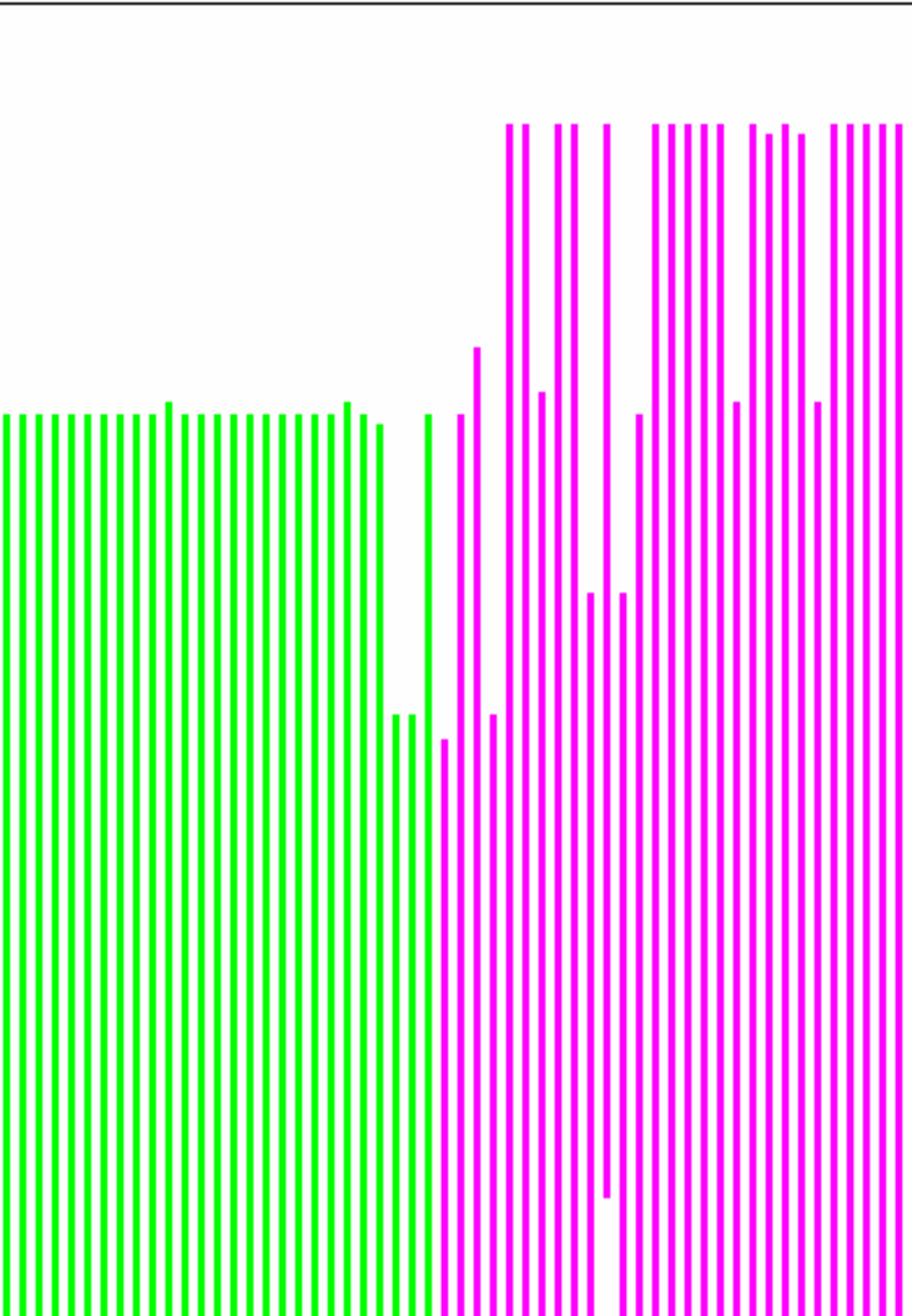
>=200

1_9468

0

50

100



Sequence detail

```
>gi|18858895|ref|NP\_571131.1| L preproinsulin [Danio rerio]
gi|9973509|sp|073727|INS\_BRARE Insulin precursor
gi|2935704|gb|AAC41261.1| L preproinsulin [Danio rerio]
Length = 108
```

```
Score = 194 bits (494), Expect = 2e-49
Identities = 96/108 (88%), Positives = 97/108 (89%)
```

```
Query: 1 MAUVMIQAGAXXXXXXXXXXXXXTNPGTPQHL CGSHLVDALYLUCGPTGFFYNPKRDVEPLLG 60
          MAUV+QAGA TNPGTPQHL CGSHLVDALYLUCGPTGFFYNPKRDVEPLLG
Sbjct: 1 MAUVMQAGALLULLVUSVUSTNPGTPQHL CGSHLVDALYLUCGPTGFFYNPKRDVEPLLG 60

Query: 61 FLPPKSAQETEVDFAFKDHAEL IRKRGIVEQCCHKPCSIFELQNYCN 108
          FLPPKSAQETEVDFAFKDHAEL IRKRGIVEQCCHKPCSIFELQNYCN
Sbjct: 61 FLPPKSAQETEVDFAFKDHAEL IRKRGIVEQCCHKPCSIFELQNYCN 108
```

```
>gi|124588|sp|P01313|INS\_CRIL0 INSULIN PRECURSOR
gi|2137094|pir||I48166 insulin precursor - golden hamster
gi|305360|gb|AAA37089.1| preproinsulin
Length = 110
```

```
Score = 75.5 bits (184), Expect = 1e-13
Identities = 45/90 (50%), Positives = 56/90 (62%), Gaps = 15/90 (16%)
```

```
Query: 27 QHLCGSHLVDALYLUCGPTGFFYNPK--RDVEPLLGLFPPKSAQETEVDFAFKDHAELI 84
          QHLCGSHLV+ALYLUCG GFFY PK R VE P+ AQ E+ D + +
Sbjct: 28 QHLCGSHLVEALYLUCGERGFFYTPKSRRGVE-----DPQVAQ-LELGGGPGADDLQTL 80

Query: 85 -----RKRGRIVEQCCHKPCSIFELQNYCN 108
          +KRGIV+QCC CS+++L+NYCN
Sbjct: 81 ALEVAQQKRGIVDQCCTSICSLYQLENYCN 110
```



Identify evolutionarily related genomic sequences

- Homologs - Orthologs - Paralogs

Annotate reference sequence

- Genic sequences - Repetitive elements - cpG islands

Align genomic sequences

- Global alignment program - Local alignment program

Identify conserved sequences

- Percent identity and length thresholds

Visualize conserved sequences

- Moving average point plot (VISTA)
- Gap-free segment plot (PipMaker)

VISTA Organization

SERVERS

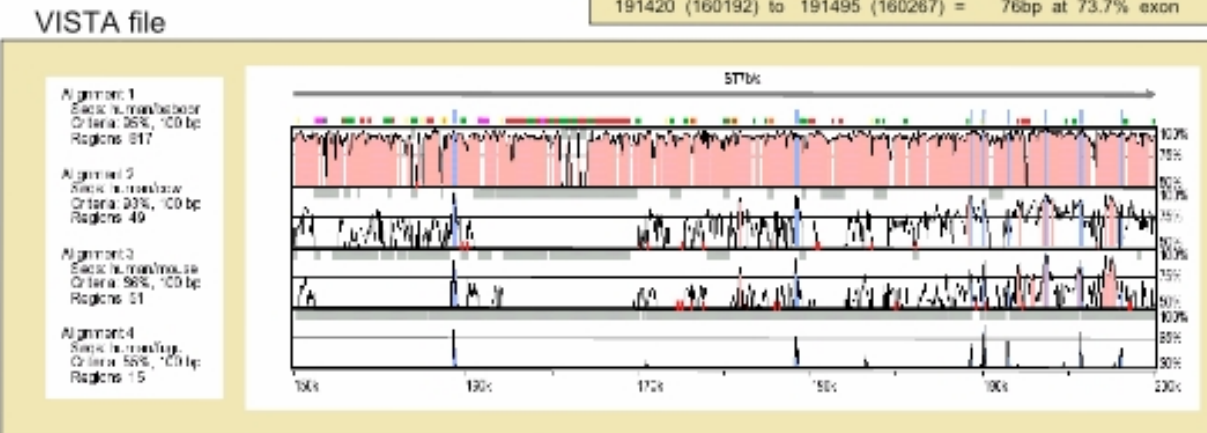
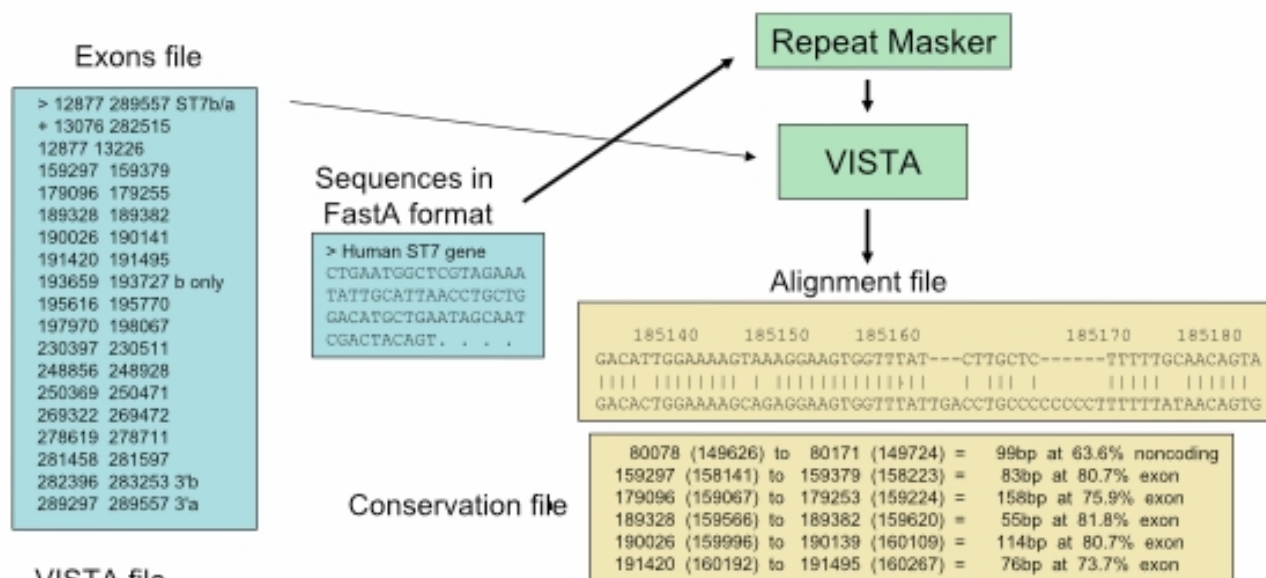
2+ orthologous sequences
Pairwise or Multiple

<http://www-gsd.lbl.gov/VISTA/>

Genome Vista
1+ to compare to
whole human or mouse

[http://pipeline.lbl.gov /](http://pipeline.lbl.gov/)

Query responses



PipMaker

VISTA

Input files

DNA sequences

Annotation of the base sequence

base sequence mask file

underlay files (for any sequence)

embedded hyperlink file

Output files

Alignments in different formats (nucleotide level)

Ordered and oriented sequence relative to first sequence

The percent identity plot

VISTA plot

dot plot

Conserved sequences

analysis of exons: splice junctions,
predicted coding sequence

Length

~2mb, time limited

4 mb

Implementation

Web server and stand alone programs,
finished and draft sequences

Underlying alignment

local

global

Features to be visualized

Order and orientation of aligned sequences

CpG islands

Genes, exons, repeats, CNSs,

Gaps in both sequences

Folding @ Home

<http://folding.stanford.edu/>

- A distributed computing PF project
download & install client software
- 1-10 ns of simulation of protein and solvent
- Issues:
 - ◆ Networking (HTTP and proxies)
 - ◆ Security (corruption of data)
 - ◆ Feedback (don't waste cycles)



Folding@home

Distributed Computing

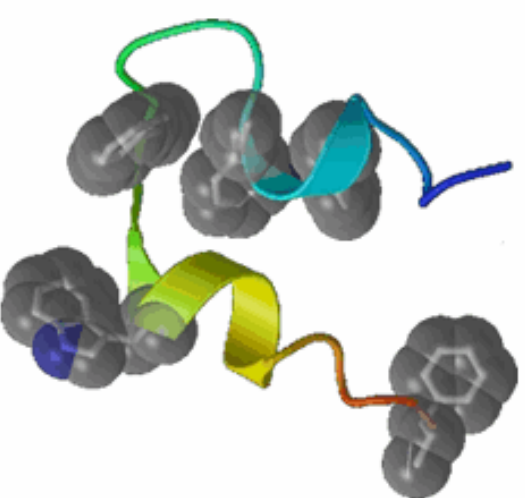
Our goal: to understand protein folding, protein aggregation, and related diseases

- [Home](#)
- [Participate \(Download\)](#)
- [Help!](#)
- [Education](#)
- [News](#)
- [Stats](#)
- [Science](#)
- [Results](#)
- [About](#)

What are **proteins** and why do they "fold"? **Proteins** are biology's workhorses -- its "**nanomachines**". Before proteins can carry out their biochemical function, they remarkably assemble themselves, or "**fold**". The process of protein folding, while critical and fundamental to virtually all of biology, remains a mystery.

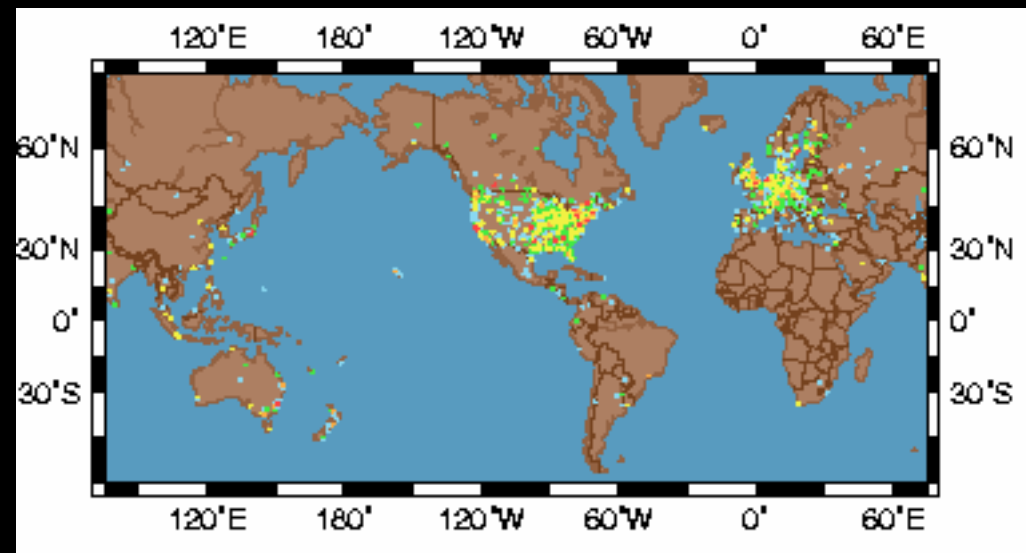
Moreover, perhaps not surprisingly, when proteins do not fold correctly (i.e. "misfold"), there can be serious effects, including many well known **diseases**, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, and Parkinson's disease.

What does Folding@Home do? Folding@Home is a distributed computing project which studies **protein folding**, misfolding, aggregation, and **related diseases**. We use novel computational methods and large scale distributed computing, to simulate timescales thousands to millions of times longer than previously achieved. This has allowed us to simulate folding for the first time, and to now direct our approach to examine folding related disease.



Results from Folding@Home simulations of villin

How can you help? You can help our project by **downloading** and running our client software. Our algorithms are designed such that for every computer that joins the project, we get a commensurate increase in simulation speed.



Summary

I. Programs for local and global alignments

PipMaker <http://bio.cse.psu.edu/>

Vista <http://sichuan.lbl.gov/vista/index.html>

Pattern Hunter <http://www.bioinformaticssolutions.com/downloads/ph-academic/>

ClustalW <http://www.ebi.ac.uk/clustalw/>

BLAST <http://www.ncbi.nlm.nih.gov/BLAST/>

LALIGN http://www.ch.embnet.org/software/LALIGN_form.html

SSEARCH <http://www.biology.wustl.edu/gcg/ssearch.html>

BLAT <http://www.genome.ucsc.edu/cgi-bin/hgBlat?command=start>

SSAHA <http://bioinfo.sarang.net/wiki/SSAHA>

II. Databases of Genomic Sequences

NCBI <http://www.ncbi.nlm.nih.gov/>

TIGR <http://www.tigr.org/>

Sanger <http://www.sanger.ac.uk/>

Ensembl <http://www.ensembl.org/>

TAIR <http://www.arabidopsis.org/home.html>

SGD <http://genome-www.stanford.edu/Saccharomyces/>

MGD <http://www.informatics.jax.org/>

Human Genome Browser <http://www.genome.ucsc.edu/>

NISC <http://www.nisc.nih.gov/>

Rat Genome Database <http://www.rgd.mcw.edu/>

FlyBase <http://flybase.bio.indiana.edu/>

Wormbase <http://brie2.cshl.org:8081/>

ExoFish <http://www.genoscope.cns.fr/externe/tetraodon/>

III. Resources for Annotated Genomic Sequences

Human Genome Browser <http://www.genome.ucsc.edu/>

Ensembl <http://www.ensembl.org/>

NCBI <http://www.ncbi.nlm.nih.gov/>

MGD <http://www.informatics.jax.org/>

FlyBase <http://flybase.bio.indiana.edu/>

Gene Annotation/Prediction Programs

GENSCAN <http://genes.mit.edu/GENSCAN.html>

GenomeScan

Sim4 <http://pbil.univ-lyon1.fr/sim4.html>

EST_Genome <http://www.sanger.ac.uk/Software/Alfresco/download.shtml>

FGENESH <http://genomic.sanger.ac.uk/gf.html>

GrailEXP <http://compbio.ornl.gov/grailexp/>

TwinScan <http://genes.cs.wustl.edu/query.html>

Genie http://www.fruitfly.org/seq_tools/genie.html

SGP <http://kiwi.ice.mpg.de/sgp-1/>

IV. Databases for homology searches

NCBI <http://www.ncbi.nlm.nih.gov/>

TIGR <http://www.tigr.org/>

MGD <http://www.informatics.jax.org/>

Ensembl <http://www.ensembl.org/>

Human Genome Browser <http://www.genome.ucsc.edu/>

SGD <http://genome-www.stanford.edu/Saccharomyces/>

Conclusion

- Ultimately, the only way to familiarize yourself with these resources is to go to the various web sites and start exploring some of the links.
- Good tutorials are available on line.

References

- Modern Genetic Analysis: Integrating Genes and Genomes

<http://bcs.whfreeman.com/mga2e/>

>Exploring Genomes: Web-Based Bioinformatics Tutorials

- Baxevanis, Andreas D.; B.F. Francis Ouellette. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. 2nd edition. 2001: John Wiley & Sons, Inc.

Acknowledgements

- Inna Dubchak
LBL: Life Sciences Division –
Genome Sciences Dubchak Lab
- Teresa Head-Gordon
UCB: Dept. of Bioengineering