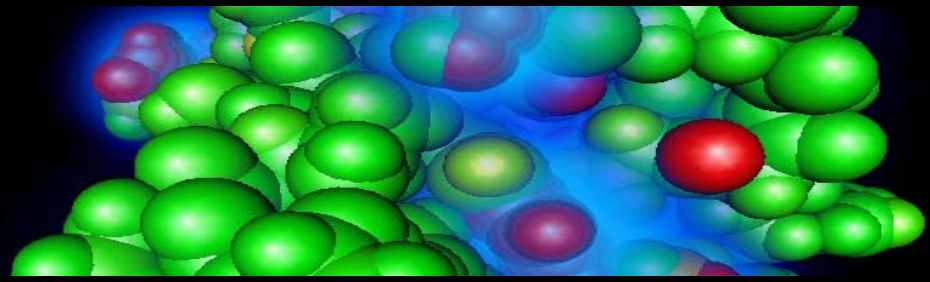


Finding the Unknown in a Sea of Data: Leveraging Human Intuition with Scientific Data Analysis

February 17, 2006

E. Wes Bethel

with help from Friends at

Lawrence Berkeley National Laboratory

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Front Matter

- **What is Berkeley Lab?**
  - General science laboratory located in Berkeley, CA run by UC for DOE.
  - Founded in 1931 as home for Cyclotron.
  - Emphasis on multidisciplinary science teams.
  - More information: www.lbl.gov

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Presentation Message

- Modern science is dominated (limited?) by information management challenges.
- Human intuition can and should play a key role in accelerating understanding in data-intensive scientific research.
- Two examples where scientific understanding is accelerated by leveraging human intuition.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Problem Statement

We live in an information dominant age.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Problem Statement

- **Information management is a limiting factor in many sciences and endeavors:**
  - Time: You have 20 minutes between tokomak experiments to analyze results from previous run and set parameters for next one.
    - Did the magnetic field lines stabilize in the last run?
    - What happened in that other experiment?

Problem Statement

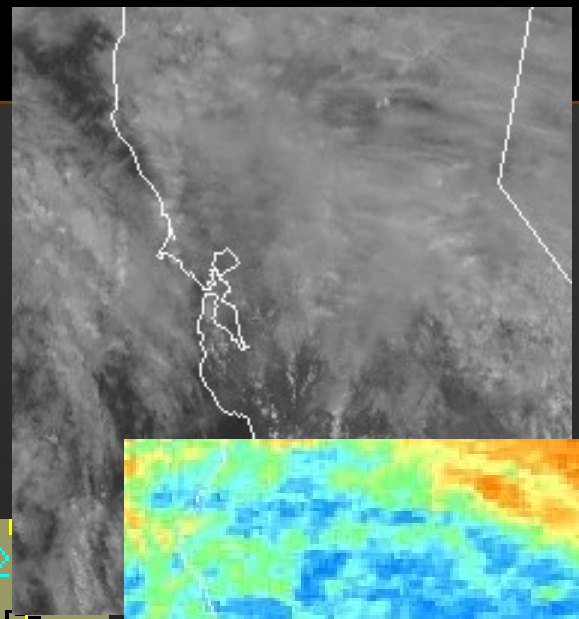
- **Simple questions give rise to startling complexity.**
  - Will a new malaria vaccine be effective?

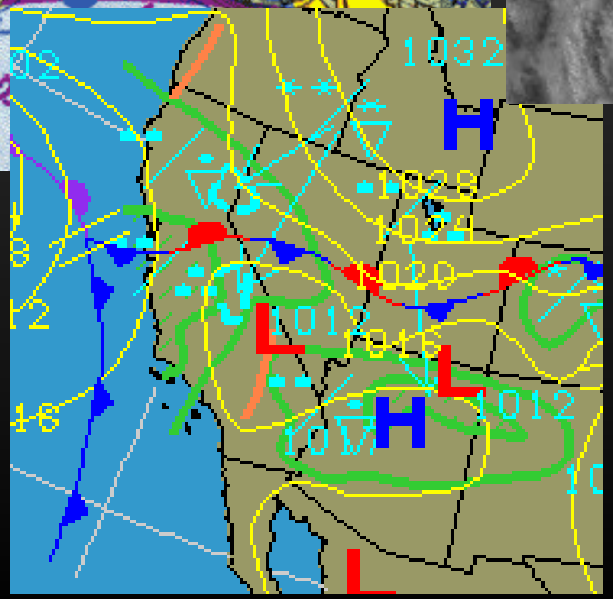
    Genome dbase, metabolic pathway dbases, prioritization, compare against human genes.
  - What is a flame front?
  - Should I fly today? (When should we launch the shuttle or schedule a landing?)

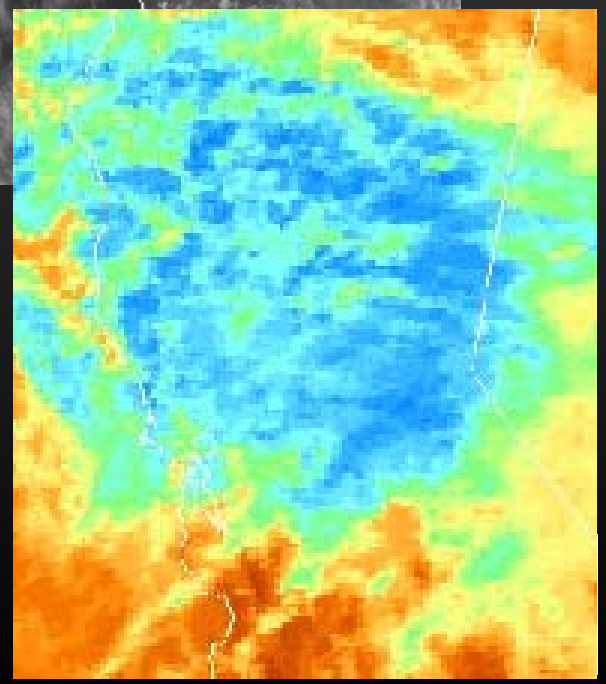
Office of Science
U.S. DEPARTMENT OF ENERGY

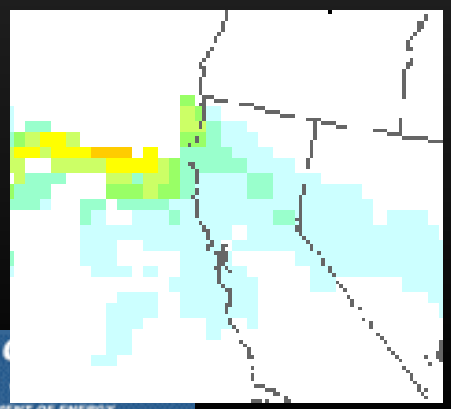
BERKELEY LAB

A Simple Question: Should I Fly Today?



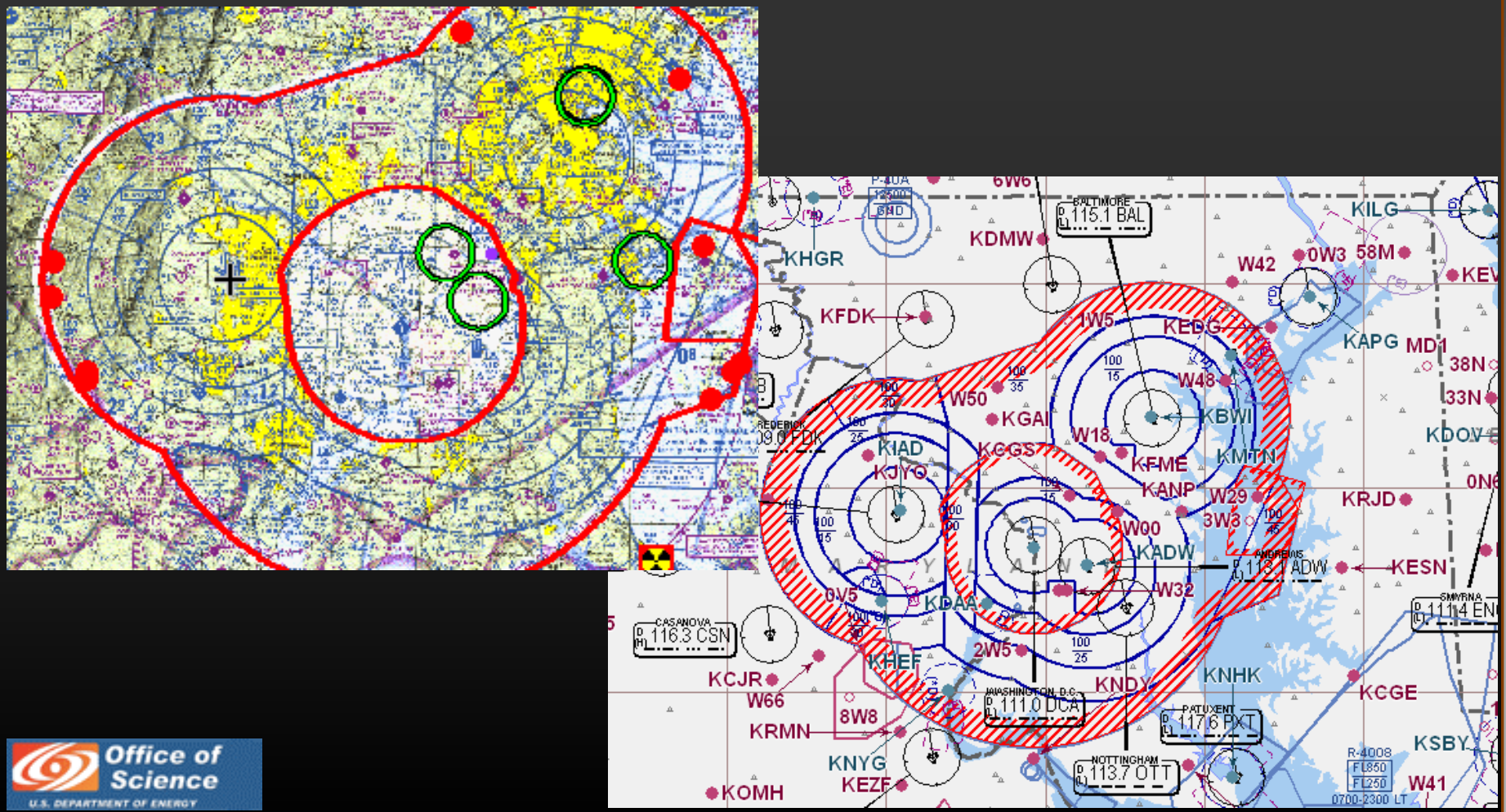






U.S. DEPARTMENT OF ENERGY

The Simple Question Becomes More Complex When Considering All Available Data



Office of Science
U.S. DEPARTMENT OF ENERGY

Dimensions of the Problem

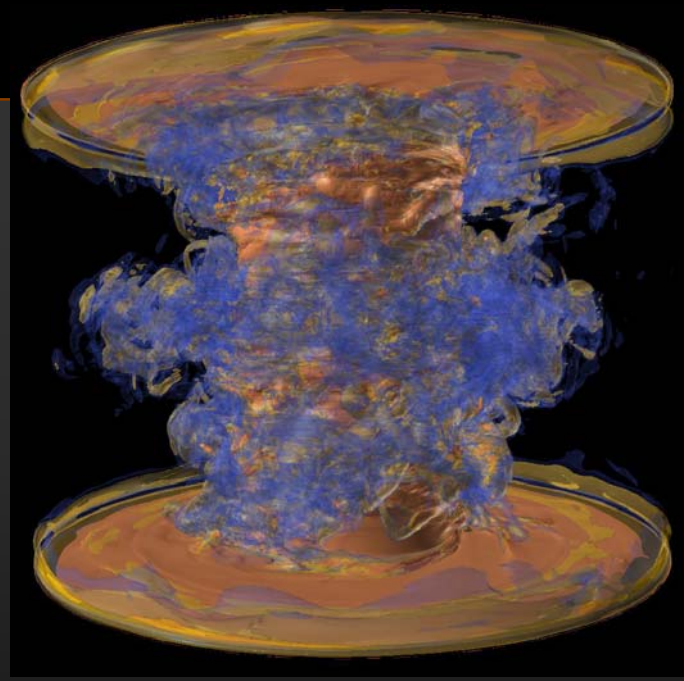
- **Data size and complexity.**
  - Where to store it? How to access it?
  - “I’m spending nearly all my time, finding, processing, organizing, and moving data—and it’s going to get much worse.”
- **N-body problem.**
  - Multiple research groups within one discipline.
  - Migration of data between disciplines.
- **Other problems: metadata management, workflows, federated data, distributed data, data analysis, …**

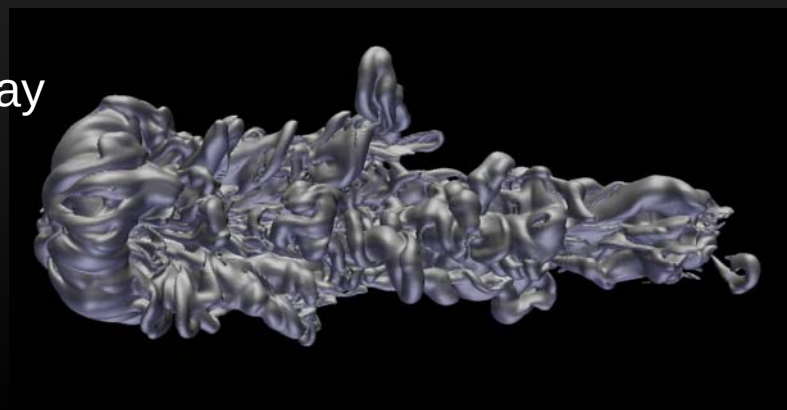
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

One “Bigger Data” Solution: Use A Bigger Hammer

- **Scalable solutions for processing larger data using existing algorithms.**
  - Faster computers, scalable tools produce increased capacity – humans ought to be able to visually process the increased load.
- **Some known problems:**
  - Doesn’t really solve the “overwhelmed with data” problem.
  - Increasing the amount of visible data may result in *less* comprehension.

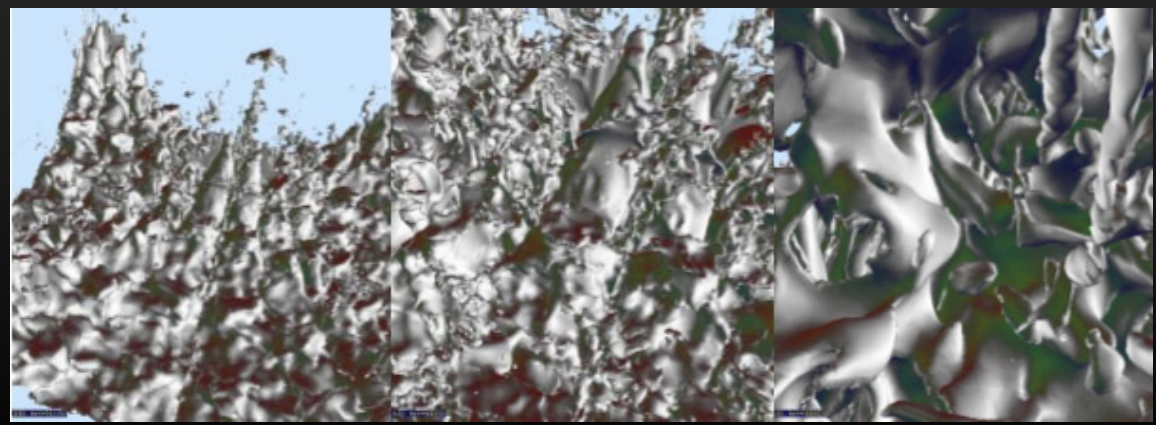




Office of Science
U.S. DEPARTMENT OF ENERGY

Another “Big Data” Solution: Save and Analyze only Interesting Data

- A researcher is focusing effort on a specific line of inquiry. Engineering vs. scientific discovery.
- Large, parallel simulation includes some visualization processing code.
- “Throwing away data” has an opportunity cost.



(Image from ASCI TSB project)

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Alternative: Query-Driven Analysis

- Combines scientific data management and visualization/analysis technology.
- Quickly locate scientifically interesting or relevant data from a larger, complete collection (don’t throw data away).
- Limit processing in downstream analysis pipeline to smaller-sized data subset.
- This approach adaptable to many different deployment alternatives: big hammer, specialized hammer, etc.

Query-Driven Visualization and Analysis

- **New capability: Bitmap Indices – find data records/cells that meet search criteria.**
  - (500<temp<1000) && (pressure<10.0mb) && (CH4>10ppm)
- **New capability: For spatial data, generate connected regions from records/cells returned by search.**
- **Exceptional performance:**
  - Searches evaluated in linear time proportional to number of hits as opposed to number of data records/points.
- **Widely applicable: Search results are input to visualization or analysis tools.**

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

What is a Bitmap Index?

Data values	b_0	b_1	b_2	b_3	b_4	b_5
0	1	0	0	0	0	0
1	0	1	0	0	0	0
5	0	0	0	0	0	1
3	0	0	0	1	0	0
1	0	1	0	0	0	0
2	0	0	1	0	0	0
0	1	0	0	0	0	0
4	0	0	0	0	1	0
1	0	1	0	0	0	0
	=0	=1	=2	=3	=4	=5

- **Compact: one bit per distinct value per object.**
- **Easy to build: faster than common B-tree**
- **Efficient to query: use bitwise logical operations.**
  - (A < 2) AND (b_0 OR b_1)
- **Efficient for multi-dimensional queries.**
  - Use bitwise operations to combine the partial results
- **What about floating point data?**

Office of Science U.S. Department of Energy

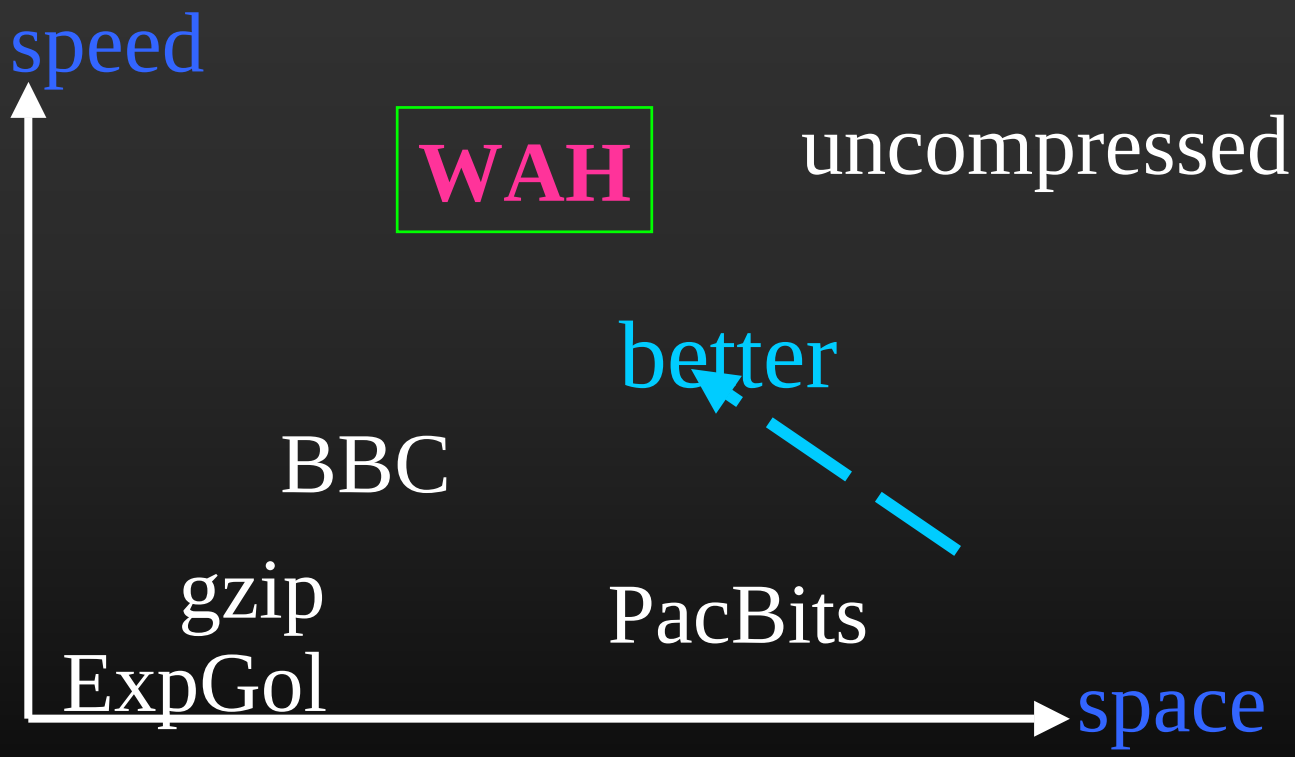
Berkeley Lab

Bitmap Index Compression

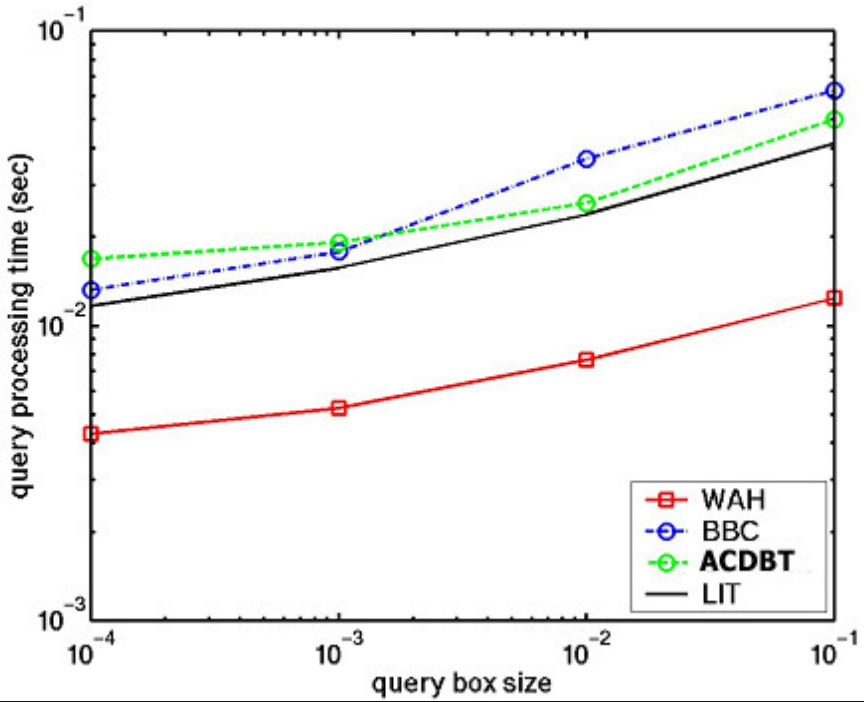
- Let N denote the number of objects and H denote the number of hits of a condition
- Using uncompressed bitmap indices, search time is O(N)
- With a good compression scheme, the search time is O(H) – the theoretical optimum.

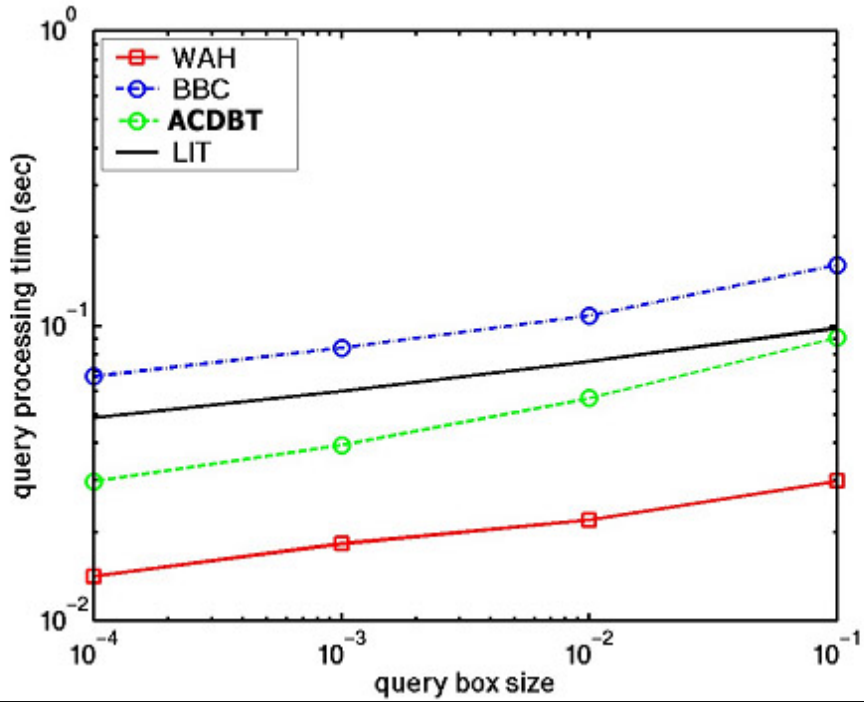
- In the worst case (completely random data), the bitmap index requires costs about 2x in data size.
- On the average, we've seen a cost of 1/10th the size of the original data.

Word-Aligned Hybrid Codes – Fast and Compact



WAH Query Performance





Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

What Does This All Mean for Scientific Research?

- **More productive science:**
  - E.g.; Locate regions of data relevant to line of scientific inquiry and focus processing/analysis on “interesting regions.”
- **Through new analysis capabilities:**
  - Traditional visualization tools (slice, crop, isosurface) fall short of meeting current scientific needs.
  - Multidimensional queries directly addresses many types of scientific inquiry.
- **With less time-to-solution:**
  - Bitmap index searches are theoretically optimum.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Some Potential Uses

Multidimensional "Data Google"

- Not only data values, but relationships between data elements.

- Scientific: physics, astronomy, biology, ….
- Economic: Credit risk assessment, ….
- Cybersecurity: internet traffic analysis, ….

Office of Science
U.S. DEPARTMENT OF ENERGY

Query-Driven Analysis Themes

- **Human judgment guides how to extract meaningful data from large and complex data collections.**
- **QDVA, when combined with interactive analysis pipelines, accommodates well-known cognitive processes:**
  - Switching between macro and micro views.
  - Data equivalent of motion parallax.
- **A patented, highly efficient data analysis capability.**

Query-Driven Analysis Future

- Multiresolution queries, temporal queries.
- Queries across federated sources.
- “Embedded” bitmap indexing as a filter in real-time, stream-processing applications.
- As the basis for comparative and integrative visual data analysis.

(Next, Predicting Protein Structure)

Predicting Protein Structure

- **Grand challenge in computational biology:**
  - Function follows from form:
    - Hemoglobin's shape allows it to carry oxygen.
    - Collagen's shape is ideal for connective tissue.
- **Why predict protein structure?**
  - Knowing shape is critical for designing therapeutic drugs.
  - Crystallization/x-ray diffraction is time-consuming, expensive and not always possible.
  - NMR methods don't work well with large proteins.
  - The simulate vs. experiment argument.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Predicting Protein Structure – Workflow

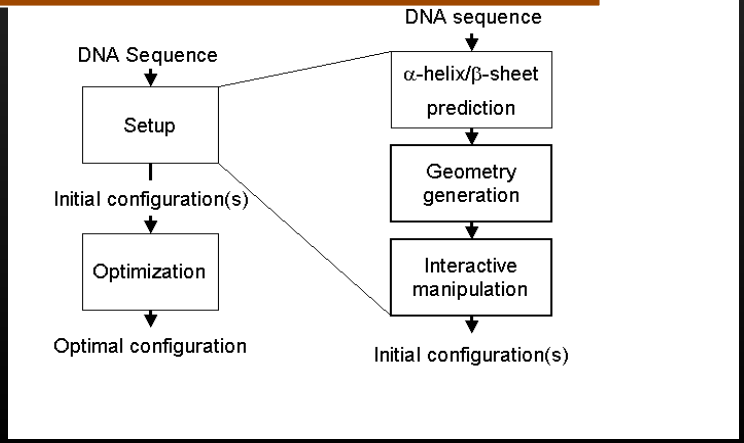
Given: an amino acid sequence,

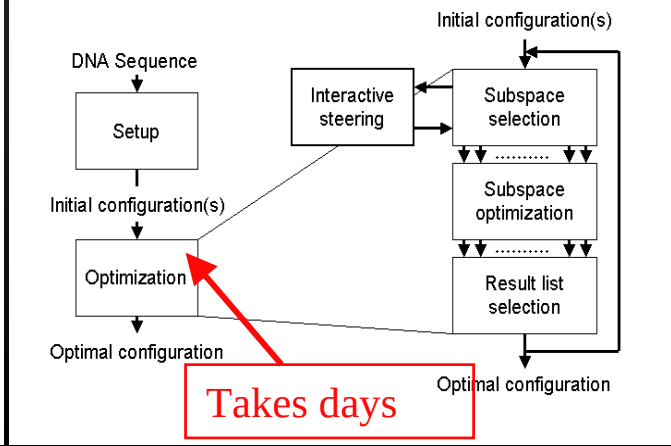
Find: an optimal protein conformation.

Many approaches are possible.

Physics-based, knowledge-based, …

```
Conf: 99999999999999999999999999999999
Pred: HHHHHHHCCCEEEEEEECCCEEEEEEECCCCCC
AA:   FKQYANDNGVDGVWTYDDATKTFTVTEMVTEVPVA
```





Office of Science
U.S. DEPARTMENT OF ENERGY

Predicting Protein Structure – Physics-Based Approach

- Given sequence of amino acids, the primary structure,
- Classify groups of them into secondary structures (alpha helices, beta strands),
- Arrange secondary structures into some 3D arrangement (tertiary structure).
- Adjust dihedral angles of atoms comprising protein chain.
- When minimal energy level is discovered, you're done!

- Problem: finding minimum energy conformation can take days on large supercomputers.

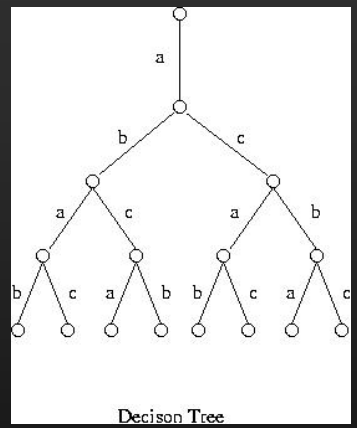
Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Predicting Protein Structure – Energy Optimization

"Find Optimal Conformation"

- **Problem: what is the minimal-energy structure of a sequence of amino acids?**
- **How to compute:**
  1. Begin with molecule in some configuration.
  2. Compute total internal energy for that configuration.
  3. Is energy minimal? If so, stop.
  4. If not, adjust dihedral angles of molecule to create a new conformation.
  5. Proceed to step 2 and repeat.



Decison Tree

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Energy Optimization Significance and Complexity

- **Theory – a protein's "final" shape is one in which it has minimum internal energy.**
  - 1972 Nobel Prize in Chemistry, Anfinsen, NIH.
- **Confirmed by comparing theoretical and experimental results.**
- **Energy Optimization is not known to be solvable:**
  - Combinatorial optimization problem.
  - Because search space is not discrete (dihedral angles), the complexity is (theoretically) not bounded by problem size.

Office of Science
U.S. DEPARTMENT OF ENERGY

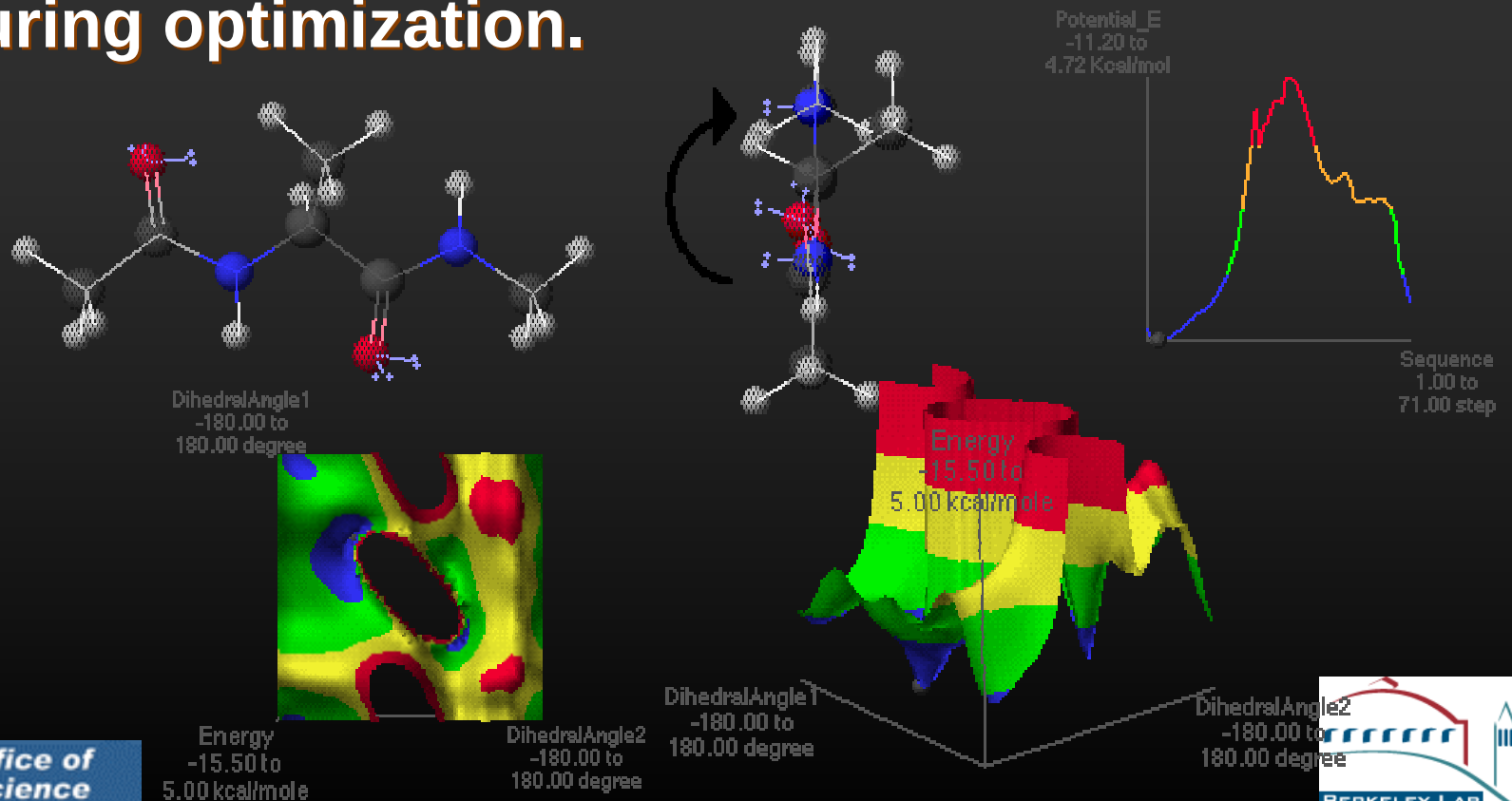
BERKELEY LAB

Our Approach: ProteinShop

- Reduce size of search space by focusing processing “most promising” candidates.
- Use human intuition to define the group of most promising candidates.
- Use human intervention during optimization to further refine the group of “most promising candidates.”

What is a Good Initial Candidate?

- One that won't "get stuck" in a local minimum during optimization.

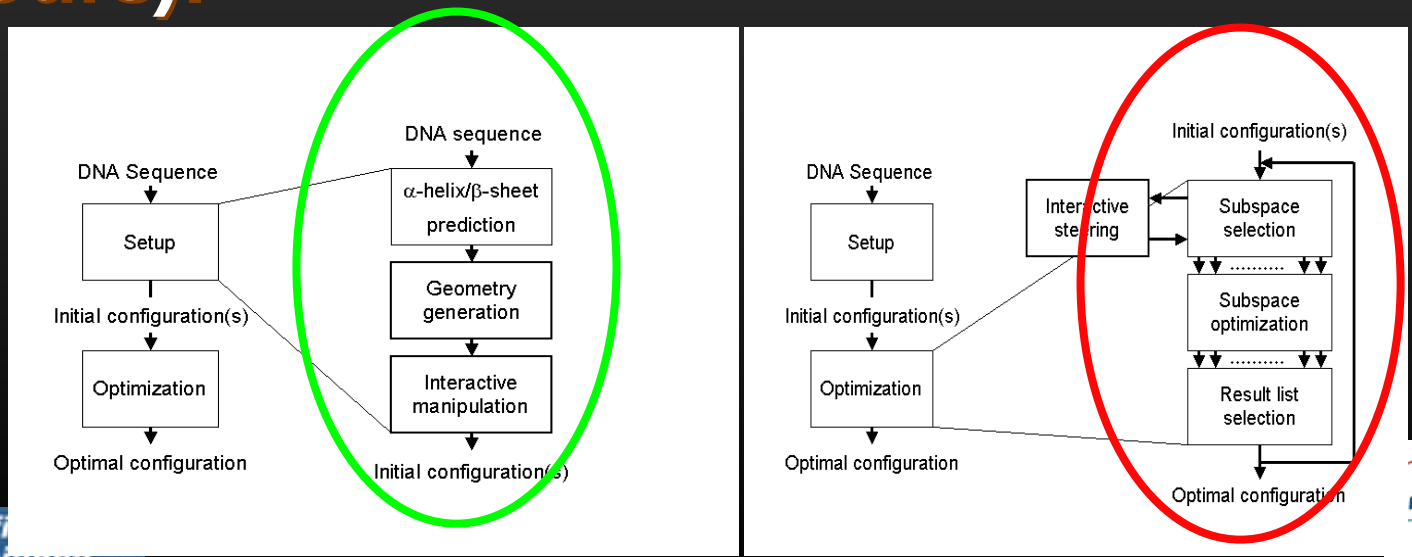


Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Good Initial Configurations Reduce Time to Solution

- Better **initial configurations** (minutes, hours) reduce time-to-solution in **optimization phase** (from days or weeks to hours).



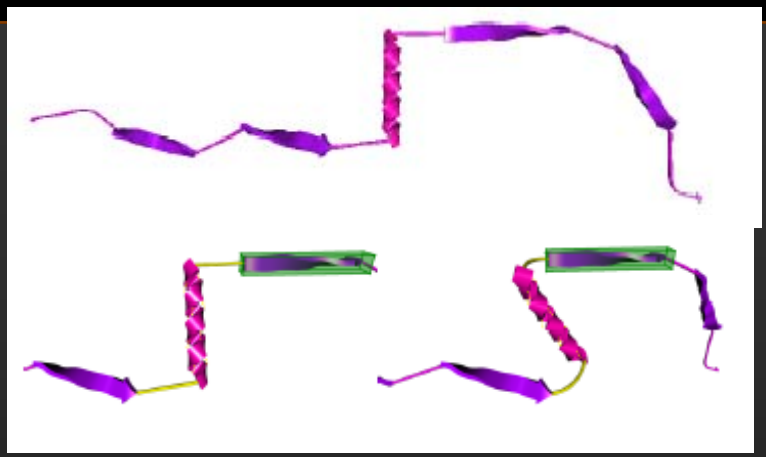
Why Domain-Specific Visualization and Interaction is Crucial

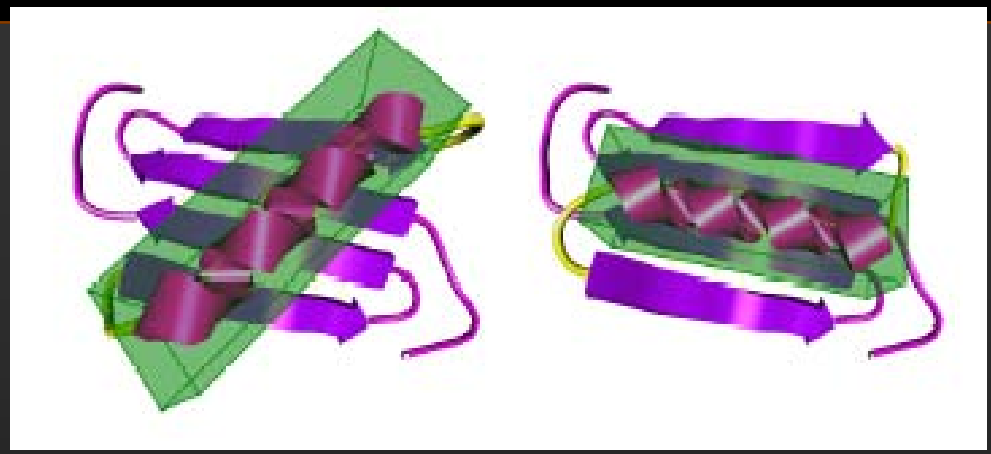
- Without constraints, it is possible to generate configurations that are not chemically meaningful (images produced with PyMol).

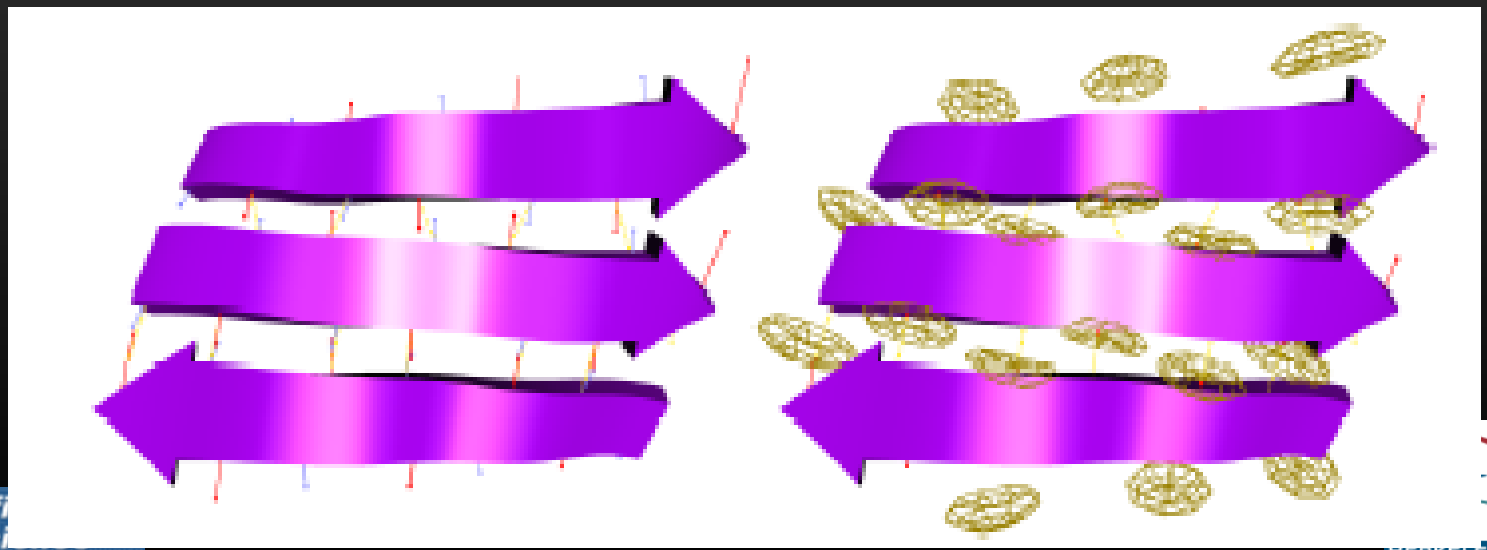
ProteinShop Interactive Manipulation

- **Inverse kinematics algorithm (from robotics) used to constrain transforms.**
  - While not chemically valid, it is physically valid, and has proven to be quite close to a chemically valid solution.
- **Interaction "handles" are secondary structures familiar to biochemists:**
  - Alpha helices, beta strands and coils.
  - Each secondary structure behaves differently during transformation.
  - Substantial improvement over other molecular visualization and manipulation tools.
- **"Guides" to help with "protein sculpting."**

ProteinShop Interactive Manipulation

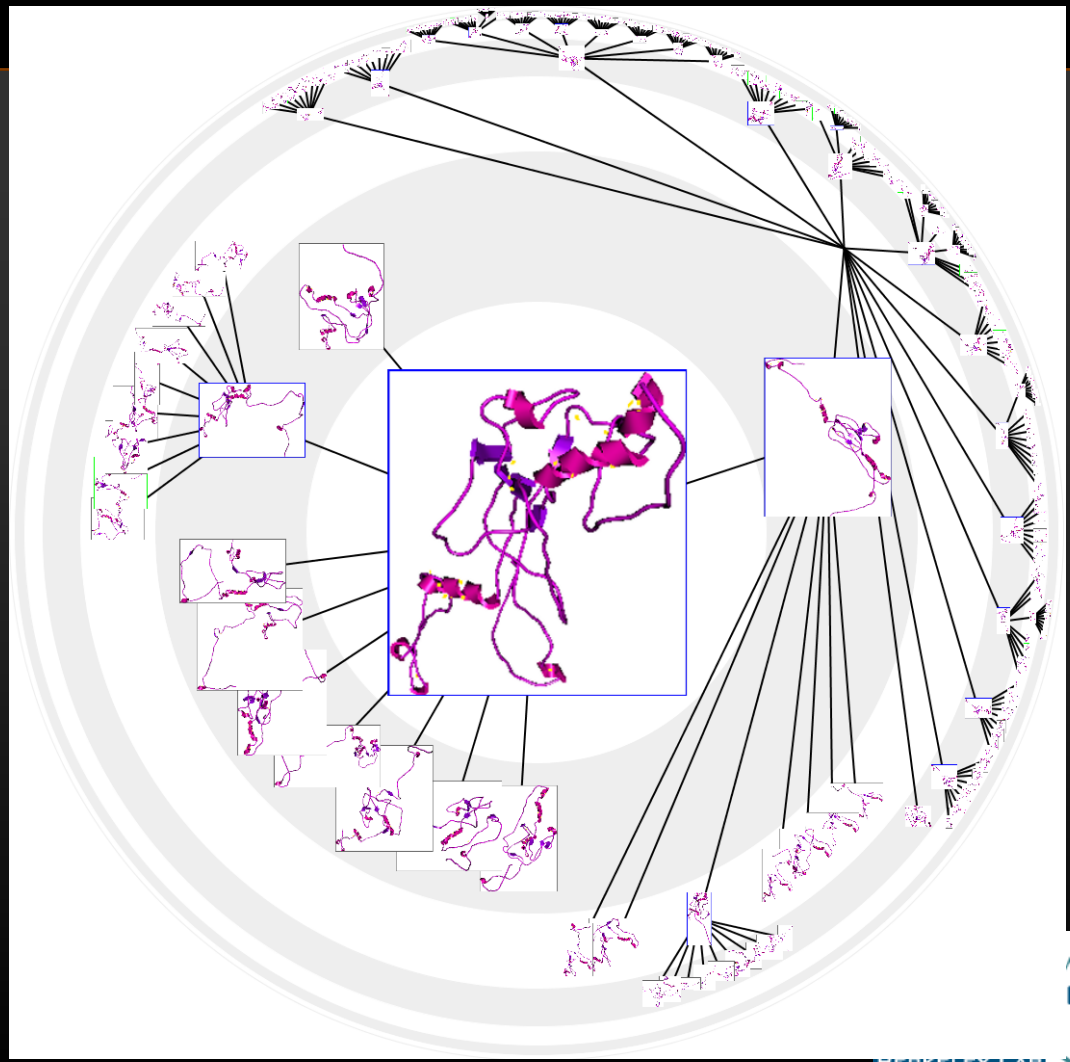




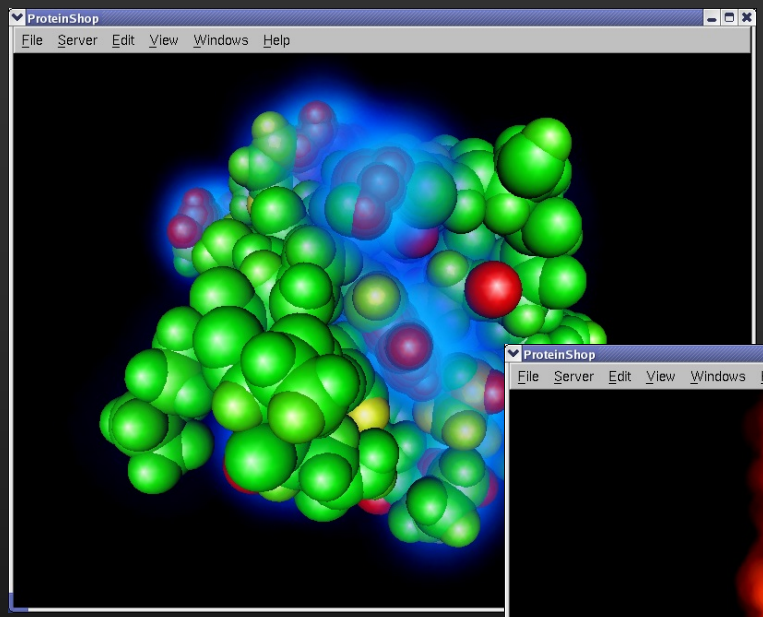


ProteinShop Guided Optimization

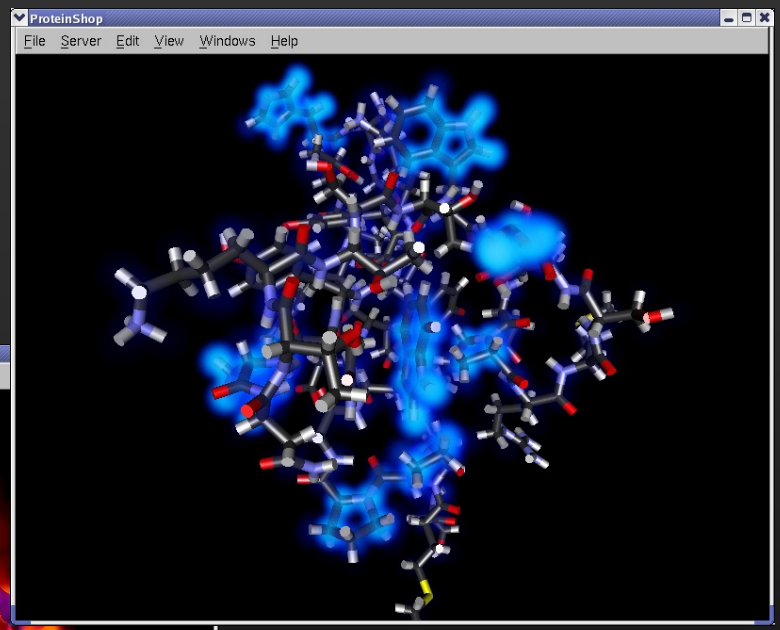
- Initial configurations used as “seed points” for optimization.
- Intermediate results – the “search tree” – is displayed for inspection.
- A human may intervene in the optimization and prune the search space.

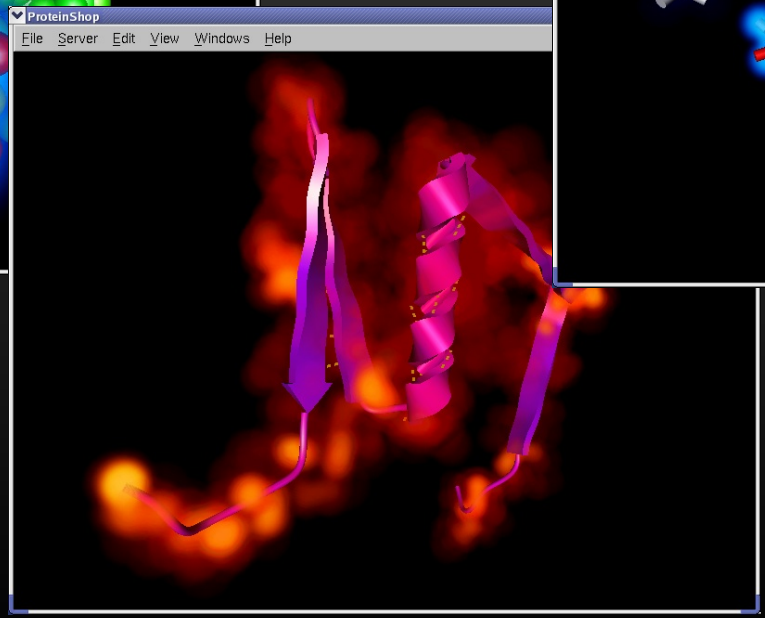


Energy Visualization (Preliminary Results)



Movie





Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ProteinShop Results (need final data)

- **CASP4 (2000, before ProteinShop)**
  - Eight targets, max length 240 amino acids.
- **CASP5 (2002, w/ProteinShop)**
  - 20 targets, max target length 417 amino acids.
  - Time to generate initial configurations reduced from days to hours.
- **CASP6 (2004, w/ProteinShop)**
  - Avg. 80th percentile in results (very good!).
  - Best score for most difficult target.

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Human Intervention Accelerates Protein Structure Prediction

- **Use of biochemical knowledge with domain-specific interactive visual analysis tools to produce good initial configurations.**
  - Reduces time to solution by avoiding local minima.
  - More computational time spent on best candidates.
- **Prune search space during optimization to reject unpromising families of optimization targets.**

Conclusions

- Modern science is dominated (limited?) by information management challenges.
- Human intuition can and should play a key role in accelerating understanding in data-intensive scientific research.
- The two examples we present offer glimpses into promising avenues of better using human knowledge to accelerate scientific discovery.

Conclusions

- Query-driven visual data analysis offers new capabilities to scientific researchers aimed at helping reduce “information overload” and information management.
- Human intuition coupled with domain-specific tools accelerate protein structure prediction by reducing time-to-solution and enabling study of larger and more complex problems.

The End

Further Reading

- NSF/NIH Workshop on Biological Data Management. http://pueblo.lbl.gov/~olken/wdmbio/ Feb 2003.
- DOE/OS Data Management Challenge. Sep 2004. http://www-user.slac.stanford.edu/rmount/dm-workshop-04/Final-Report-Work-Area/
- Improved Searching for Spatial Features in Spatio-Temporal Data <http://sdm.lbl.gov/~kurts/research/region-growing-lbnl-sept2004.ps>, Technical Report, LBNL-56376, Berkeley, California, September 2004
- LBNL Visualization Group: http://vis.lbl.gov/

Office of Science U.S. Department of Energy

Berkeley Lab

Further Reading, ctd.

- ProteinShop. http://proteinshop.lbl.gov/. Fall 2004.
- Unfolding Proteins, Krell Institute. http://www.krellinst.org/csgf/mag/2003/print.cgi?id=2100
- The Art of Protein Structure Prediction. http://www.eurekalert.org/features/doe/2004-12/ddoe-tao122204.php
- "Interactive Protein Manipulation." In Proceedings of *IEEE Visualization 2003*, October 19-24, 2003, Seattle, Washington, USA, pp 581-588. LBNL-52414 (Best Application Paper Award) http://vis.lbl.gov/Publications/2003/Kreylos-ProteinShop-Vis2003.pdf

Office of Science U.S. Department of Energy

Berkeley Lab