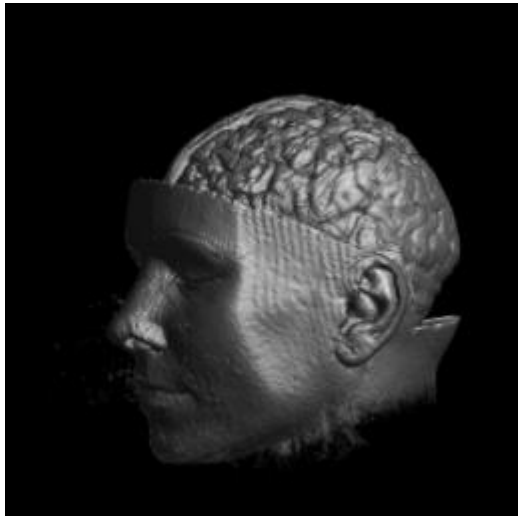
High Performance Computing Applications in Biology

Ananth Grama
Department of Computer Sciences
Purdue University
http://www.cs.purdue.edu/people/ayg
ayg@cs.purdue.edu

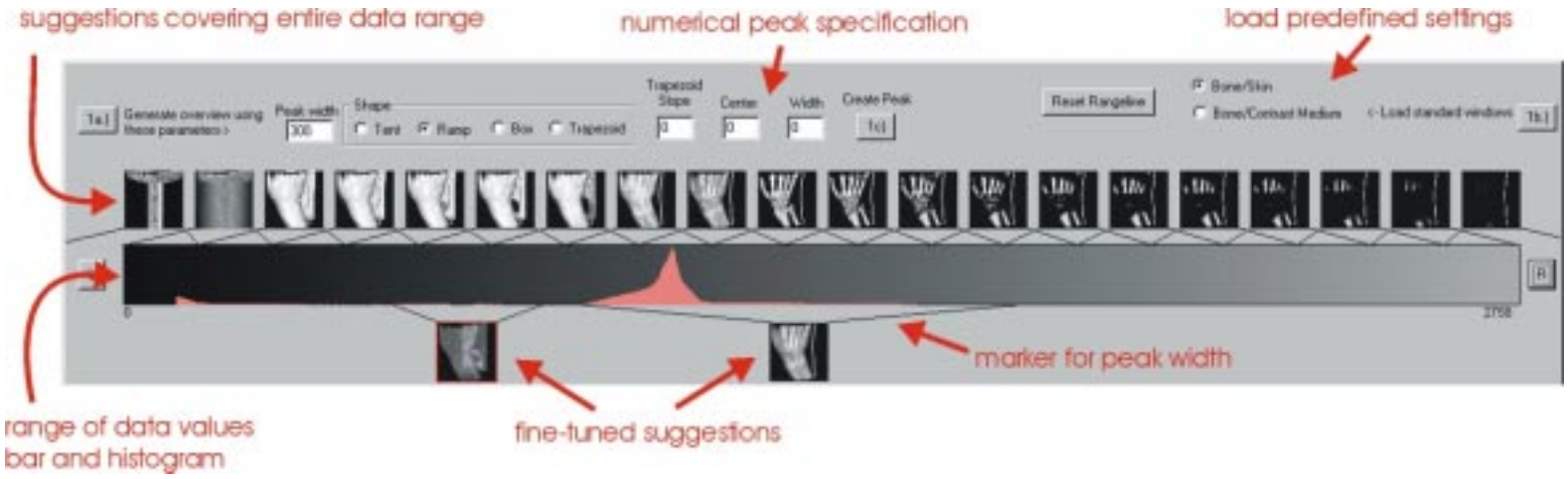
Part I: Some Success Stories Modeling, Visualization, and Analysis.

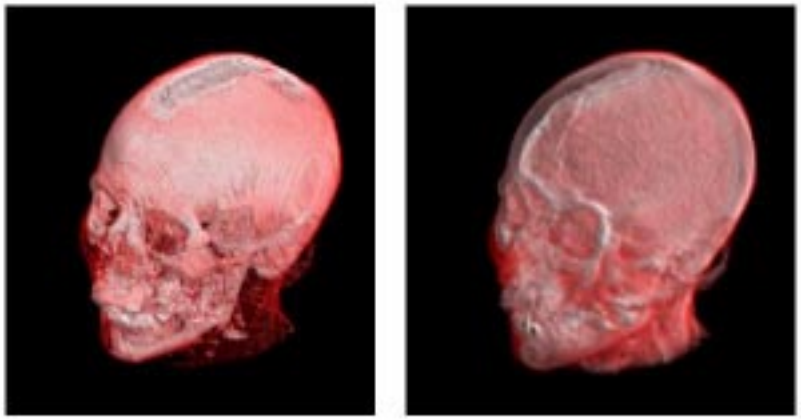
Imaging, Reconstruction, and Analysis

- Computerized Tomography
- Magnetic Resonance Imaging



Volume Visualization





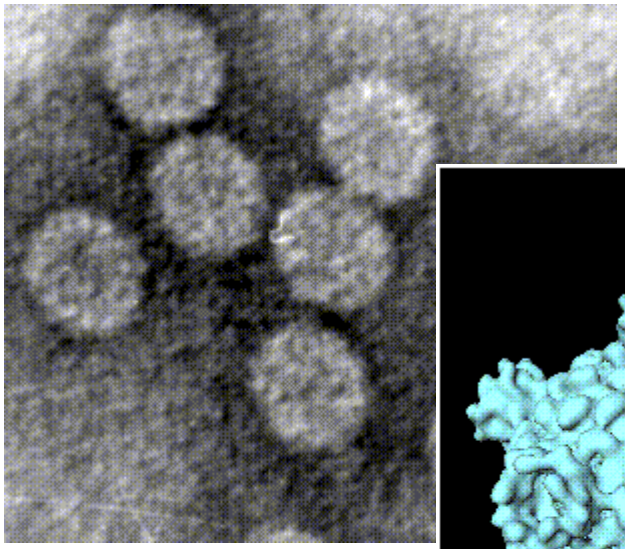
CT Head. DVR of decimated (left) and original (right) data sets

(Visualization and animation group, Technical University of Vienna)

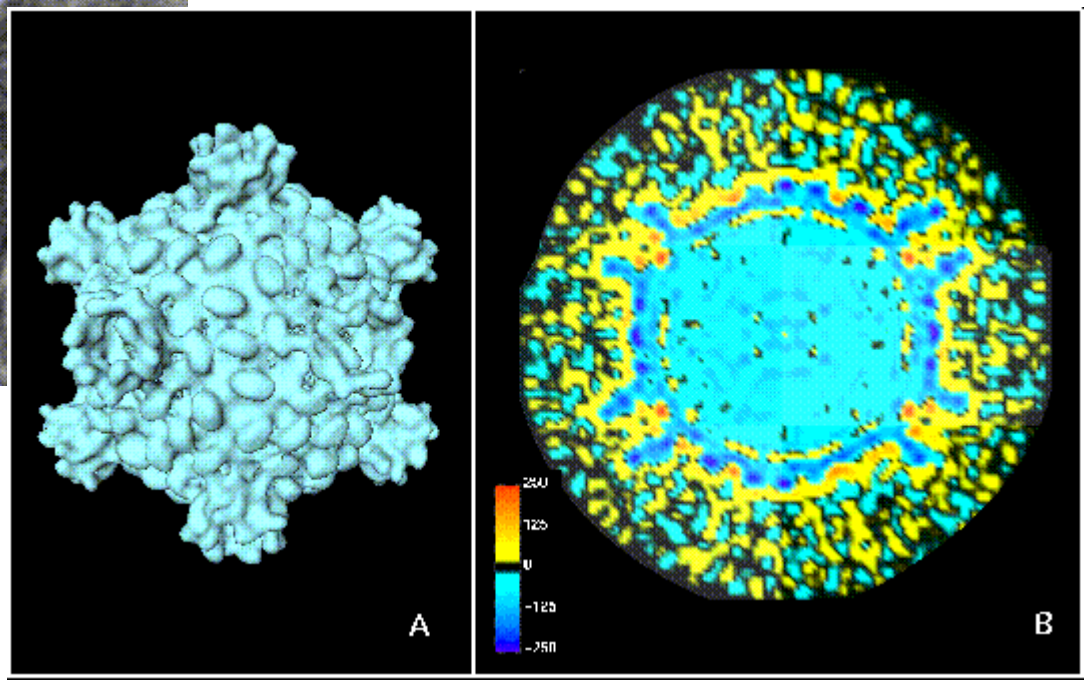
Reconstruction and Inverse Problems

- Given excitation and response, compute the structure that results in observed response (reconstruction).
- Given response and structure, compute excitation to achieve desired effect (non-invasive cauterization of tumors, focused electromagnetic fields).

Reconstruction (Virus Structure)

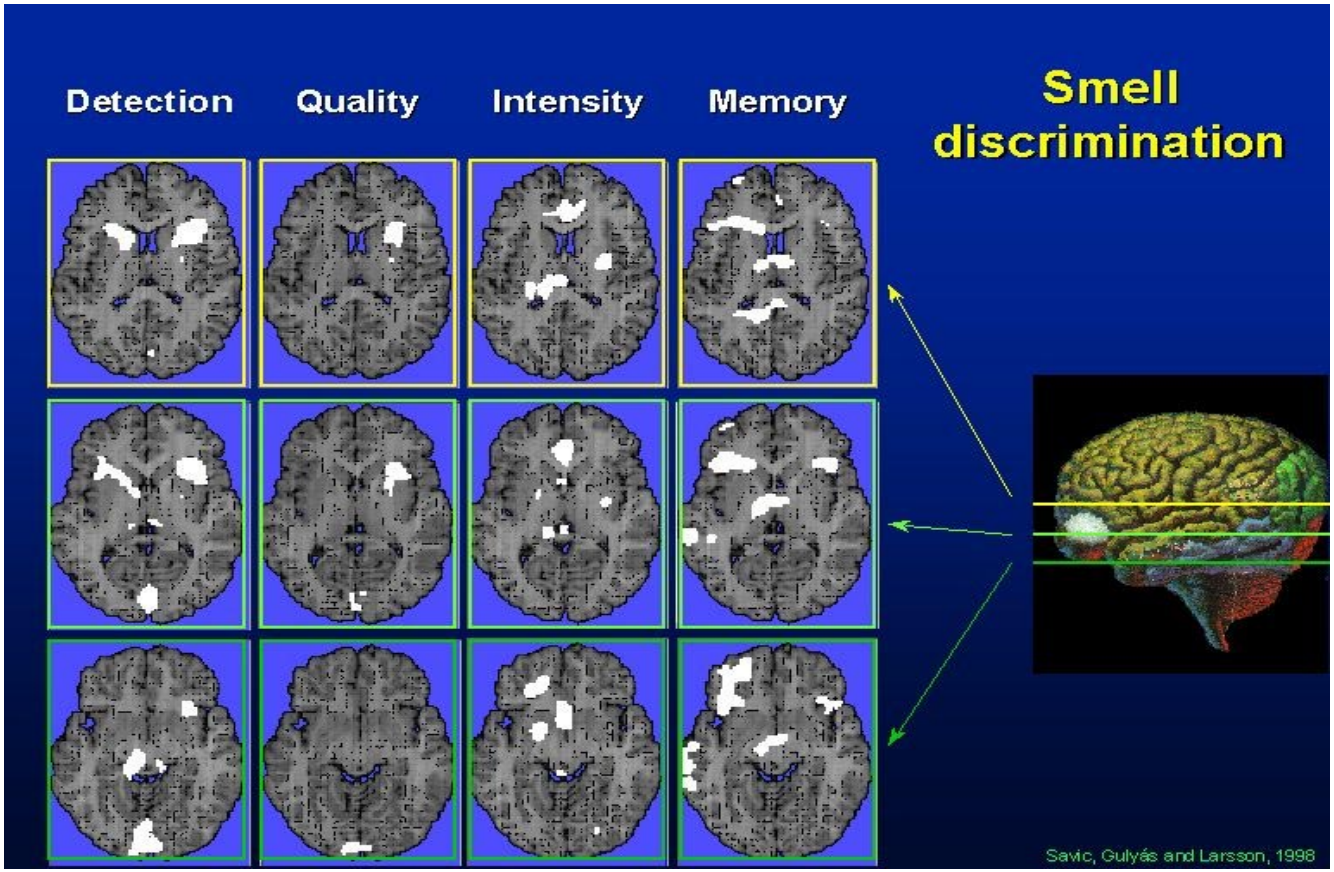


Digitized Electron Micrograph

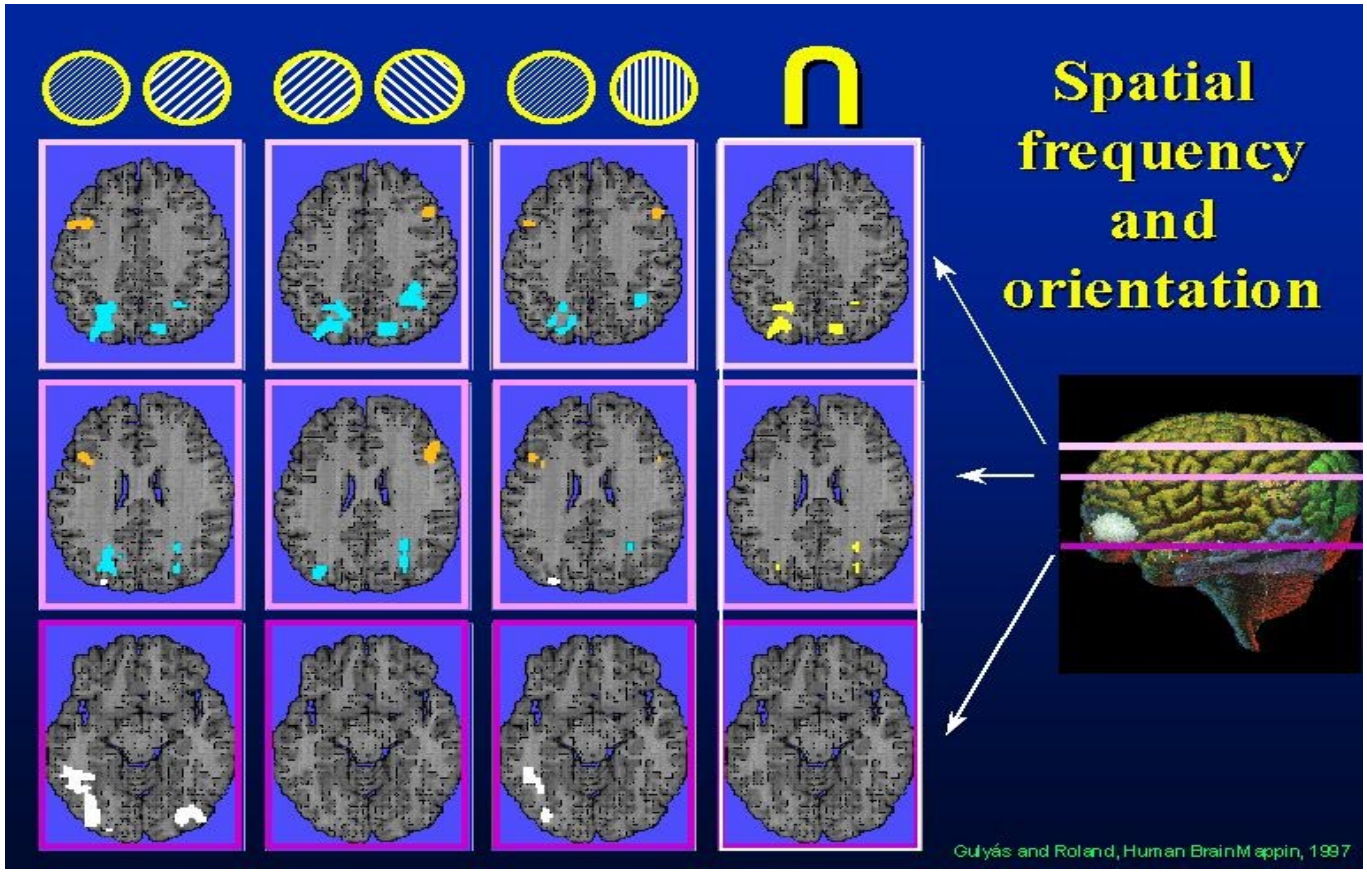


Reconstructed isosurface rendering and cross-section of Mammalian Reovirus core reconstructed from micrographs.

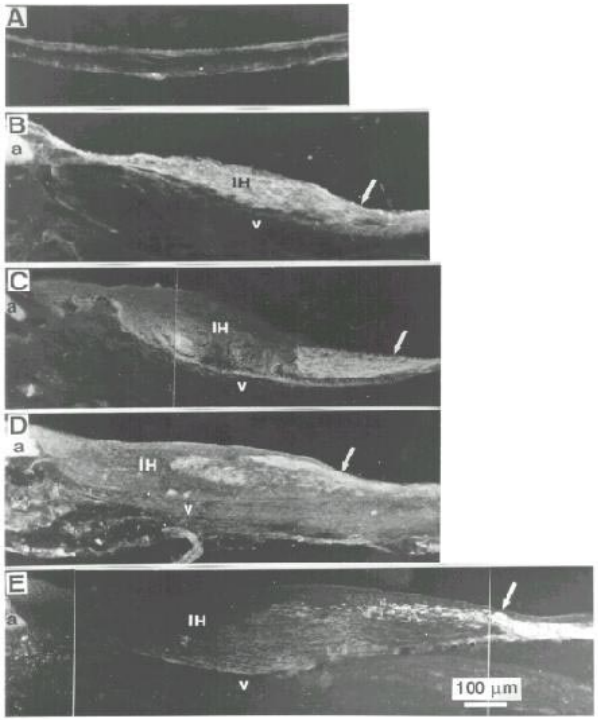
Functional Mapping of the Brain



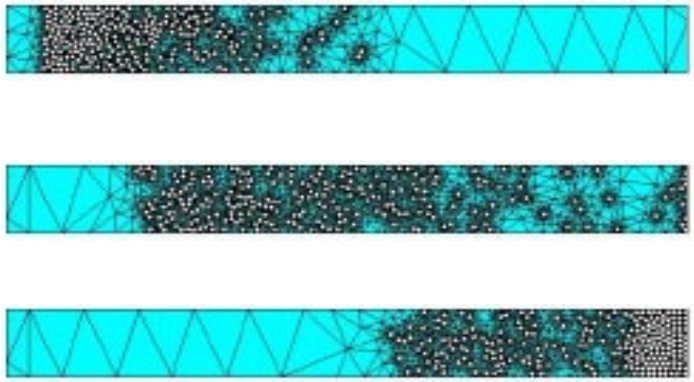
Functional Mapping of the Brain



Modeling and Simulation (Blood flows through vein grafts)

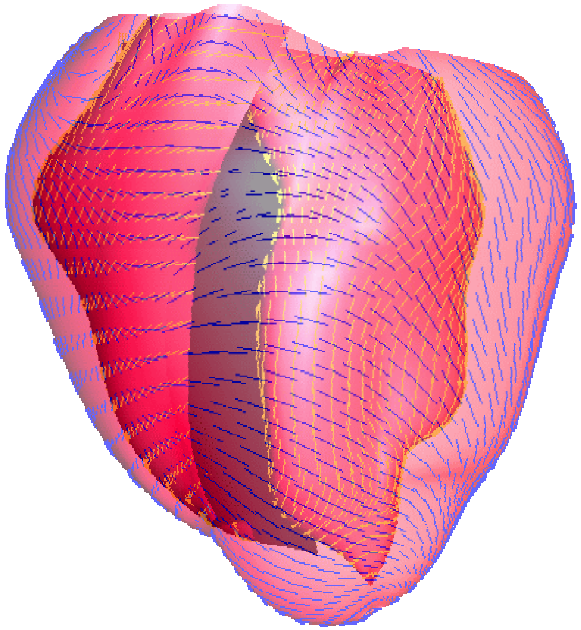


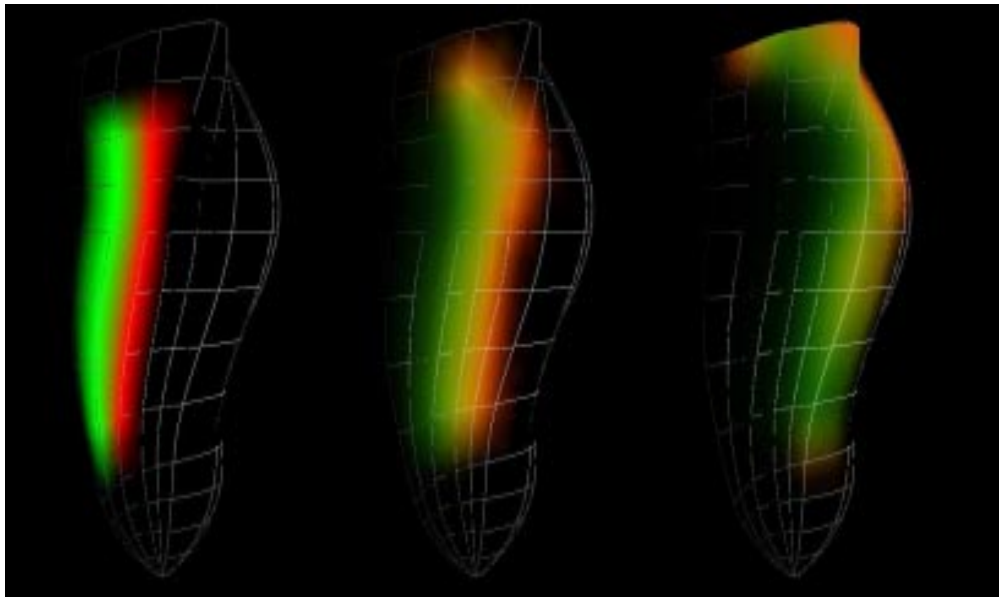
Micrographs showing progressive growth of intimal hyperplasia in the proximal region of the vein graft on (A) day 0; (B) day 5; (C) day 10; (D) day 20; and (E) day 30.



Simulation of blood as an incompressible particulate fluid using finite element analysis

Modeling and Simulation (Impulse Propagation)





Finite element model of a rabbit heart and simulated 2D propagation of electrical activation wave on the left ventricular inner surface. Fully excited tissue is shown in red and refractory tissue in green (14 ms intervals) [Fred Vetter, SDSC].

Part II: Computational Genomics and Proteomics

Building Blocks of Life

- The gene is the basic unit of heredity
- Composed of DNA, genes carry the imprint that describes the appearance and behavior
- The DNA in a gene is expressed by first being transcripted to mRNA
- This message is translated to form amino-acid sequences that are the building blocks of proteins
- Proteins are responsible for various functions/manifestations

Proteins and Genes

- One can argue for direct correlation between genetic structure and medical conditions.
- However, in conditions such as cancer, a combination of several genetic alterations are necessary.
- Detecting such networks of gene expressions (epigenetics) is an extremely difficult task.

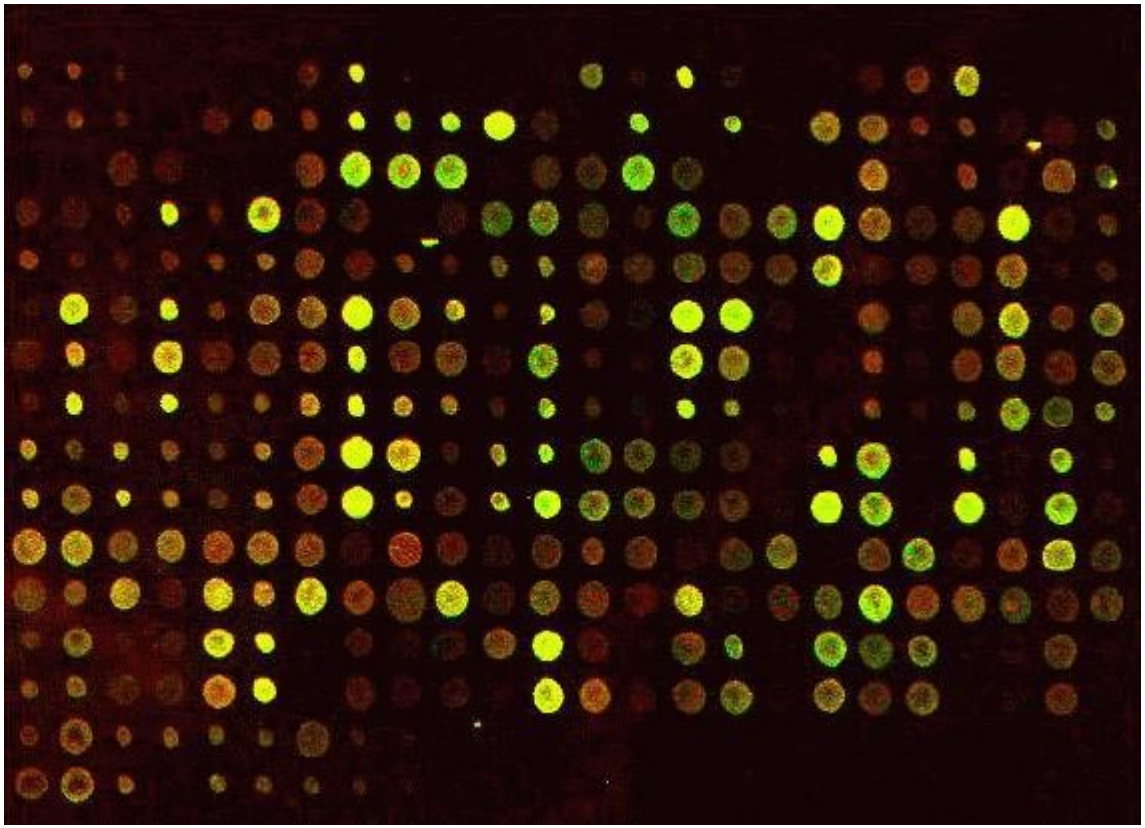
Analyzing Gene Expression - Microarray Analysis

- With advances in high-density DNA microarray technology, it has become possible to screen large numbers of genes to see whether or not they are active under various conditions.
- This is gene-expression profiling, and there has been an expectation that it will revolutionize diagnosis of various conditions.

Microarray Analysis - Cancer Diagnosis

- Tumor behavior is dictated by the expression of thousands of genes.
- Micro-array analysis allows this behavior and the clinical consequences to be predicted.
- For example, using clustering analysis, Alizadeh et al. separate diffuse large B-cell lymphoma (DLBCL) into two categories, which had marked differences in overall survival of the patients concerned (Nature, 2/3/2000).

Microarray Analysis



Biological variation of gene expression in Loblolly Pine cones by microarray analysis. Genes that are co-expressed under similar stress/drought conditions are clustered to hypothesize and/or update gene regulatory systems (Alscher and Heath, 2000).

Proteins: Structure and Function

- Amino-acids form the building blocks of proteins.
- There are roughly 20 natural and about 80 modified amino-acids.
- Proteins contain upwards of several thousand amino-acids in polypeptide bonds.
- A very large combinatorial space is thus available for assembling proteins.

Proteins: Some Facts and Figures

- Human body makes between 50,000 and 100,000 proteins.
- Proteins typically survive in the body for about two days and are dismantled and or discarded.
- Amino-acids have an 8-atom body and a side-chain of between 1 and 18 atoms.
- All side chains contain hydrogen, most contain carbon, many contain oxygen, and some contain nitrogen and sulphur.

Why Study Proteins?

- In theory, it is possible to directly correlate genes to associated activity (as opposed to going via the protein).
- Disrupt selected genes and study its effect.
- However, disrupting a single gene invarably impacts expression of many other genes, making direct causality difficult to establish.
- Recent studies have also shown lack of correlation between mRNA and associated protein in a given cell at a given point of time.

Studying Proteins - Structure

- The amino-acid sequence formed from mRNA folds up in a matter of seconds to form a 3D structure.
- This is the functional protein - it interacts with other molecules (lock and key mechanisms) to regulate body function.
- The challenge is to determine this 3D structure (and subsequently function) from amino-acid sequences, which are easy to obtain.

Computing Protein Structure

- Given a sequence (FASTA):

```
>CG2B_MARGL
MLNGENVDSRIMGKVATRASSKGVKSTLGTRGALENISNVARNNLQAGAK
KELVKAKRGMTKSKATSSLQSVMGLNVEPMEKAKPQSPEPMDMSEINSAL
EAFSQNLLEGVEDIDKNDFDNPQLCSEFVNDIYQYMRKLEREFKVRTDYM
TIQEITERMRSILIDWLVQVHLRFHLLQETLFLTIQILDRYLEVQPVSKN
KLQLVGVTSMLIAAKYEEMYPPEIGDFVYITDNAYTKAQIRSMECNILRR
LDFSLGKPLCIHFLRRNSKAGGVDGQKHTMAKYLMELTLPEYAFVPYDPS
EIAAAALCLSSKILEPDMEWGTTLVHYSAYSEDHLMPIVQKMALVLKNAP
TAKFQAVRKKYSSAKFMNVSTISALTSSTVMDLADQMC
```

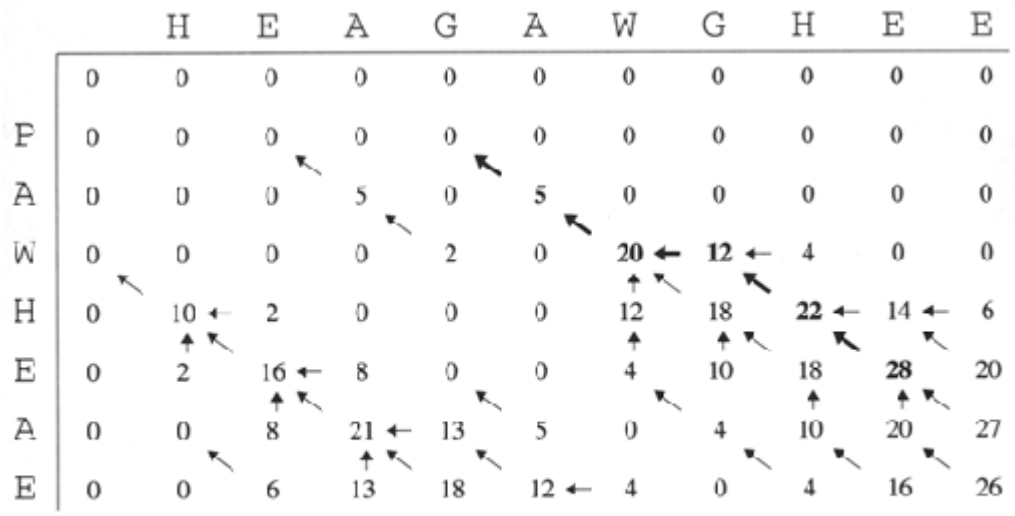
- What is its 3D structure?

A alanine	P proline
B aspartate	Q glutamine
C cystine	R arginine
D aspartate	S serine
E glutamate	T threonine
F phenylalanine	U selenocysteine
G glycine	V valine
H histidine	W tryptophan
I isoleucine	Y tyrosine
K lysine	Z glutamate
L leucine	X any
M methionine	* trans. stop
N asparagine	- gap of indeter. length

Computing Protein Structure

- Find other proteins (with known structures) with similar amino-acid sequences and piece together the structure (and infer basic function) from them.
- The most commonly used matching algorithm is due to Smith and Waterman.
- Smith-Waterman algorithm is a dynamic programming algorithm that assigns a score to each pair of bases using positive scores for related residues and negative scores for substitutions and gaps.

Smith-Waterman Example



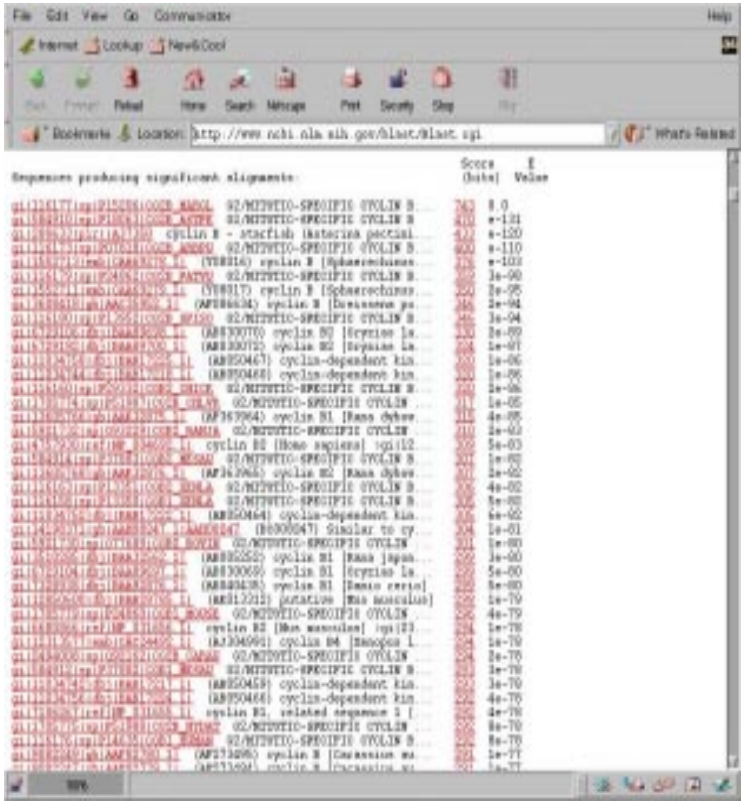
		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

AWGHE
AW-HE

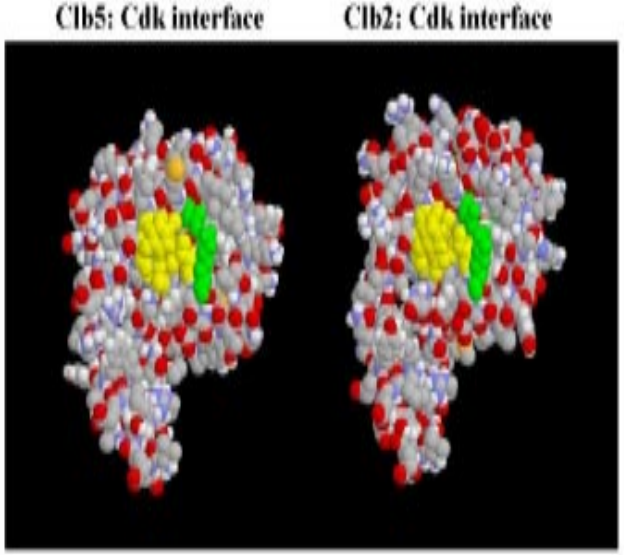
From Durbin et al. 1998

Protein Structure - Tools

- Perhaps the most popular tool is BLAST (`http://www.ncbi.nlm.nih.gov/BLAST`)



Protein Structure - BLAST



S-phase and M-phase cyclin - yellow residues make up hydrophobic interface, green side chains are glutamic acid and lysine residues.

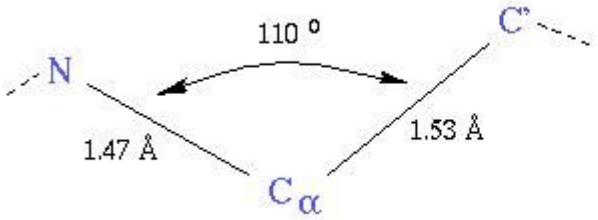
Protein Structure and Function - A Geometric Approach

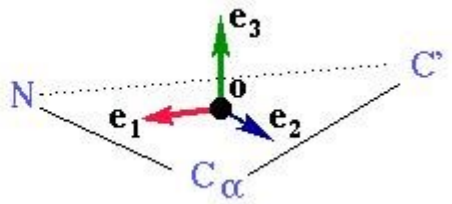
- Protein function is determined by 3D substructures, called binding motifs.
- Proteins with similar 3D motifs often exhibit similar biological properties.
- A number of algorithms for extracting 3D motifs have been proposed.

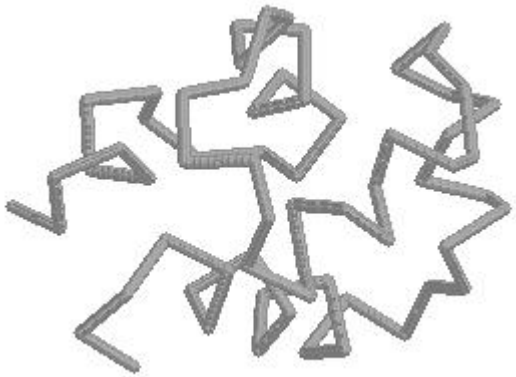
Geometric Hashing

- Pre-compute a database of redundant representations based on local features (unordered set of geometric configurations of amino-acids).
- Each amino-acid has 4 atoms participating in the backbone, 3 of them are always in the same configuration.
- Use these 3 atoms to define the configuration of the amino-acid in space.
- To find similar substructures, we now have to find two subset of frames that are in the same configuration up to a global rigid transformation.

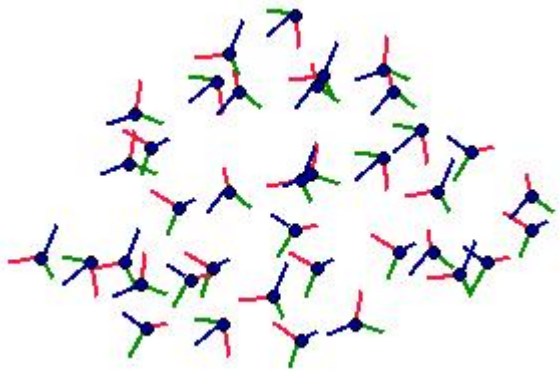
Geometric Hashing







Backbone of the Protein



Protein modeled as an unordered set of frames.

(Xavier Pennec, INRIA)

Protein Structure- An Energy-Minimization Approach

- Protein folds into a configuration of minimum energy.
- For this reason, in a water substrate, oil-loving amino-acids tend to bury themselves and water-loving amino acids tend to orient themselves at the surface.
- A simple energy model can be derived from Coloumbic and Lennart-Jones potentials.

Energy Minimization

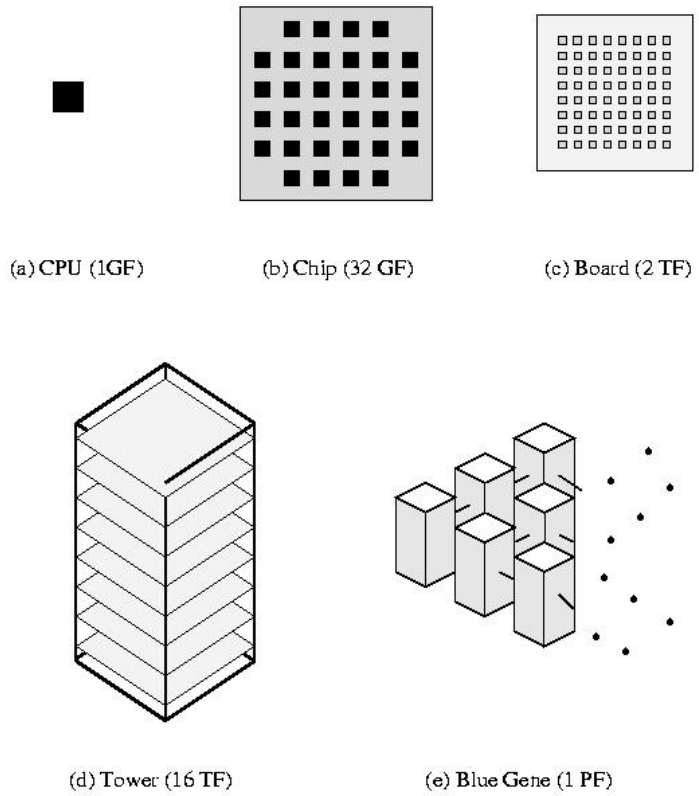
- From an initial configuration, allow atoms to move based on the potential.
- Closed form solutions are very difficult for more than 3 bodies.
- The problem here is that every atom is impacted by every other atom ($O(n^2)$ interactions per time-step for n atoms).
- For a 10,000 particle system, with a 1 femtosecond timestep, simulation of a second would take about 20 days on a petaFLOP computer.

Energy Minimization

- Algorithmic Improvements:
  - A cluster of particles far away from an observation point can be approximated by a point charge.
  - More sophisticated approximations have been developed based on multipole series.
  - Using a hierarchy of sub-domains and multipole approximation, timestep complexity can be reduced to O(n)! (Fast Multipole Method).

Energy Minimization

- Computational Improvements.



The IBM Blue Gene is designed with the molecular dynamics problem in mind and is capable of 10^{15} floating point operations per second!

Part III: Outstanding Challenges

Some Challenges in Bioinformatics

- Characterize the structure and function of bio-molecules.
- Characterize various biological pathways and the role of intermediate compounds.
- Design bio-molecules to accomplish specific function (blocking the effect of or generation of specific molecules)

Some Challenges in Modeling

- Develop accurate mathematical models for various processes.
- Develop algorithms and software using these mathematical models.
- Use the software to achieve desired physical function within the body.

Some Challenges in Infrastructure

- Handling extremely large datasets.
- Analysis (correlations, dominant and deviant patterns, etc.) and visualization of datasets.
- Modeling of complex deformable geometries.
- Inverse problems.